SYNTHETIC NUCLEIC ACID MOLECULE COMPOSITIONS AND METHODS OF PREPARATION

Statement of Government Rights

The invention was made at least in part with a grant from the Government of the United States of America (grant DMI-9402762 from the National Science Foundation). The Government has certain rights to the invention.

Background of the Invention

Transcription, the synthesis of an RNA molecule from a sequence of DNA is the first step in gene expression. Sequences which regulate DNA transcription include promoter sequences, polyadenylation signals, transcription factor binding sites and enhancer elements. A promoter is a DNA sequence capable of specific initiation of transcription and consists of three general regions. The core promoter is the sequence where the RNA polymerase and its cofactors bind to the DNA. Immediately upstream of the core promoter is the proximal promoter which contains several transcription factor binding sites that are responsible for the assembly of an activation complex that in turn recruits the polymerase complex. The distal promoter, located further upstream of the proximal promoter also contains transcription factor binding sites. Transcription termination and polyadenylation, like transcription initiation, are site specific and encoded by defined sequences. Enhancers are regulatory regions, containing multiple transcription factor binding sites, that can significantly increase the level of transcription from a responsive promoter regardless of the enhancer's orientation and distance with respect to the promoter as long as the enhancer and promoter are located within the same DNA molecule. The amount of transcript produced from a gene may also be regulated by a post-transcriptional mechanism, the most important being RNA splicing that removes intervening sequences (introns) from a primary transcript between splice donor and splice acceptor sequences.

Natural selection is the hypothesis that genotype-environment interactions occurring at the phenotypic level lead to differential reproductive success of individuals and therefore to modification of the gene pool of a population. Some properties of nucleic acid molecules that are acted upon by natural selection include codon usage frequency, RNA secondary structure, the efficiency of intron splicing, and interactions with transcription factors or other nucleic acid binding proteins. Because of the degenerate nature of the genetic code, these properties can be optimized by natural selection without altering the corresponding amino acid sequence.

Under some conditions, it is useful to synthetically alter the natural nucleotide sequence encoding a polypeptide to better adapt the polypeptide for alternative applications. A common example is to alter the codon usage frequency of a gene when it is expressed in a foreign host cell. Although redundancy in the genetic code allows amino acids to be encoded by multiple codons, different organisms favor some codons over others. It has been found that the efficiency of protein translation in a non-native host cell can be substantially increased by adjusting the codon usage frequency but maintaining the same gene product (U.S. Patent Nos. 5,096,825, 5,670,356, and 5,874,304).

However, altering codon usage may, in turn, result in the unintentional introduction into a synthetic nucleic acid molecule of inappropriate transcription regulatory sequences. This may adversely effect transcription, resulting in anomalous expression of the synthetic DNA. Anomalous expression is defined as departure from normal or expected levels of expression. For example, transcription factor binding sites located downstream from a promoter have been demonstrated to effect promoter activity (Michael et al., 1990; Lamb et al., 1998; Johnson et al., 1998; Jones et al., 1997). Additionally, it is not uncommon for an enhancer element to exert activity and result in elevated levels of DNA transcription in the absence of a promoter sequence or for the presence of transcription regulatory sequences to increase the basal levels of gene expression in the absence of a promoter sequence.

Thus, what is needed is a method for making synthetic nucleic acid molecules with altered codon usage without also introducing inappropriate or unintended transcription regulatory sequences for expression in a particular host cell.

Summary of the Invention

The invention provides a synthetic nucleic acid molecule comprising at least 300 nucleotides of a coding region for a polypeptide, having a codon composition differing at more than 25% of the codons from a wild type nucleic acid sequence encoding a polypeptide, and having at least 3-fold fewer, preferably at least 5-fold fewer, transcription regulatory sequences than would result if the differing codons were randomly selected. Preferably, the synthetic nucleic acid molecule encodes a polypeptide that has an amino acid sequence that is at least 85%, preferably 90%, and most preferably 95% or 99% identical to the amino acid sequence of the naturally-occurring (native or wild type) polypeptide (protein) from which it is derived. Thus, it is recognized that some specific amino acid changes may also be desirable to alter a particular phenotypic characteristic of the polypeptide encoded by the synthetic nucleic acid molecule. Preferably, the amino acid sequence identity is over at least 100 contiguous amino acid residues. In one embodiment of the invention, the codons in the synthetic nucleic acid molecule that differ preferably encode the same amino acids as the corresponding codons in the wild type nucleic acid sequence.

The transcription regulatory sequences which are reduced in the synthetic nucleic acid molecule include, but are not limited to, any combination of transcription factor binding sequences, intron splice sites, poly(A) addition sites, enhancer sequences and promoter sequences. Transcription regulatory sequences are well known in the art.

It is preferred that the synthetic nucleic acid molecule of the invention has a codon composition that differs from that of the wild type nucleic acid sequence at more than 30%, 35%, 40% or more than 45%, e.g., 50%, 55%, 60% or more of the codons. Preferred codons for use in the invention are those which are employed more frequently than at least one other codon for the same amino acid in a particular organism and, more preferably, are also not low-usage codons in that organism and are not low-usage codons in the organism used to clone or screen for the expression of the synthetic nucleic acid molecule (for example, *E. coli*). Moreover, preferred codons for certain amino acids (i.e., those amino acids that have three or more codons,), may include two or more codons that are employed more frequently than the other (non-preferred) codon(s). The presence of codons in the synthetic nucleic acid molecule that are employed more frequently in one organism than in another organism results in a synthetic nucleic acid molecule which, when introduced into the cells of the organism that employs those codons more frequently, is expressed in those cells at a level that is greater than the expression of the wild type or parent nucleic acid sequence in those cells. For example, the synthetic nucleic acid molecule of the invention is expressed at a level that is at least about 110%, e.g., 150%, 200%, 500% or more (1000%, 5000%, or 10000%) of that of the wild type nucleic acid sequence in a cell or cell extract under identical conditions (such as cell culture conditions, vector backbone, and the like).

In one embodiment of the invention, the codons that are different are those employed more frequently in a mammal, while in another embodiment the codons that are different are those employed more frequently in a plant. A particular type of mammal, e.g., human, may have a different set of preferred codons than another type of mammal. Likewise, a particular type of plant may have a different set of preferred codons than another type of plant. In one embodiment of the invention, the majority of the codons which differ are ones that are preferred codons in a desired host cell. Preferred codons for mammals (e.g., humans) and plants are known to the art (e.g., Wada et al., 1990). For example, preferred human codons include, but are not limited to, CGC (Arg), CTG (Leu), TCT (Ser), AGC (Ser), ACC (Thr), CCA (Pro), CCT (Pro), GCC (Ala), GGC (Gly), GTG (Val), ATC (Ile), ATT (Ile), AAG (Lys), AAC (Asn), CAG (Gln), CAC (His), GAG (Glu), GAC (Asp), TAC (Tyr), TGC (Cys) and TTC (Phe) (Wada et al., 1990). Thus, preferred "humanized" synthetic nucleic acid molecules of the invention have a codon composition which differs from a wild type nucleic acid sequence by having an increased number of the preferred human codons, e.g. CGC, CTG, TCT, AGC, ACC, CCA, CCT, GCC, GGC, GTG, ATC, ATT, AAG, AAC, CAG, CAC, GAG, GAC, TAC, TGC, TTC, or any combination thereof. For example, the synthetic nucleic acid molecule of the invention may have an increased number of CTG or TTG leucine-encoding codons, GTG or GTC valine-encoding codons, GGC or GGT glycine-encoding codons, ATC or ATT isoleucine-encoding codons, CCA or CCT proline-encoding codons, CGC or CGT arginine-encoding codons, AGC or TCT serine-encoding codons, ACC or ACT threonine-encoding codon, GCC or GCT alanine-encoding codons, or any combination thereof, relative to the wild type nucleic acid sequence. Similarly, synthetic nucleic acid molecules having an increased number of codons that are employed more frequently in plants, have a codon composition which differs from a wild type or parent nucleic acid sequence by having an increased number of the plant codons including, but not limited to, CGC (Arg), CTT (Leu), TCT (Ser), TCC (Ser), ACC (Thr), CCA (Pro), CCT (Pro), GCT (Ser), GGA (Gly), GTG (Val), ATC (Ile), ATT (Ile), AAG (Lys), AAC (Asn), CAA (Gln), CAC (His), GAG (Glu), GAC (Asp), TAC (Tyr), TGC (Cys), TTC (Phe), or any combination thereof (Murray et al., 1989). Preferred codons may differ for different types of plants (Wada et al., 1990).

The choice of codon may be influenced by many factors such as, for example, the desire to have an increased number of nucleotide substitutions or decreased number of transcription regulatory sequences. Under some circumstances (e.g. to permit removal of a transcription factor binding site) it may be desirable to replace a non-preferred codon with a codon other than a preferred codon or a codon other than the most preferred codon. Under other circumstances, for example, to prepare codon distinct versions of a synthetic nucleic acid molecule, preferred codon pairs are selected based upon the largest number of mismatched bases, as well as the criteria described above.

The presence of codons in the synthetic nucleic acid molecule that are employed more frequently in one organism than in another organism, results in a synthetic nucleic acid molecule which, when introduced into a cell of the organism that employs those codons, is expressed in that cell at a level which is greater than the level of expression of the wild type or parent nucleic acid sequence.

A synthetic nucleic acid molecule of the invention may encode a selectable marker protein or a reporter molecule. However, the invention applies to any gene and is not limited to synthetic reporter genes or synthetic selectable marker genes. In one embodiment of a synthetic nucleic acid molecule of the invention that is a reporter molecule, the synthetic nucleic acid molecule encodes a luciferase having a codon composition different than that of a wild type or parent *Renilla* luciferase or a beetle luciferase nucleic acid sequence. A synthetic click beetle luciferase nucleic acid molecule of the invention may optionally encode the amino acid valine at position 224 (i.e., it emits green light), or may optionally encode the amino acid histidine at position 224, histidine at position 247, isoleucine at position 346, glutamine at position 348 or combination thereof (i.e., it emits red light). Preferred synthetic luciferase nucleic acid molecules that are related to a wild type *Renilla* luciferase nucleic acid sequence include, but are not limited to, SEQ ID NO:21 (Rlucver2) or SEQ ID NO:22 (Rluc-final). Preferred synthetic luciferase nucleic acid molecules that are related to click beetle luciferase nucleic acid sequences include, but are not limited to, SEQ ID NO:7 (GRver5), SEQ ID NO:8 (GR6), SEQ ID NO:9 (GRver5.1), SEQ ID NO:14 (RDver5), SEQ ID NO:15 (RD7), SEQ ID NO:16 (RDver5.1), SEQ ID NO:17 (RDver5.2) or SEQ ID NO:18 (RD156-1H9).

The invention also provides an expression cassette. The expression cassette of the invention comprises a synthetic nucleic acid molecule of the invention operatively linked to a promoter that is functional in a cell. Preferred promoters are those functional in mammalian cells and those functional in plant cells. Optionally, the expression cassette may include other sequences, e.g., restriction enzyme recognition sequences and a Kozak sequence, and be a part of a larger polynucleotide molecule such as a plasmid, cosmid, artificial chromosome or vector, e.g., a viral vector.

Also provided is a host cell comprising the synthetic nucleic acid molecule of the invention, an isolated polypeptide (e.g., a fusion polypeptide encoded by the synthetic nucleic acid molecule of the invention), and compositions and kits comprising the synthetic nucleic acid molecule of the invention or the polypeptide encoded thereby in suitable container means and, optionally, instruction means. Preferred isolated polypeptides include, but are not limited to, those comprising SEQ ID NO:31 (GRver5.1), SEQ ID NO:226 (Rluc-final), or SEQ ID NO:223 (RD156-1H9).

The invention also provides a method to prepare a synthetic nucleic acid molecule of the invention by genetically altering a parent (either a wild type or another synthetic) nucleic acid sequence. The method may be used to prepare a synthetic nucleic acid molecule encoding a polypeptide comprising at least 100 amino acids. One embodiment of the invention is directed to the preparation of synthetic genes encoding reporter or selectable marker proteins. The method of the invention may be employed to alter the codon usage frequency and decrease the number of transcription regulatory sequences in any open reading frame or to decrease the number of transcription regulatory sites in a vector backbone. Preferably, the codon usage frequency in the synthetic nucleic acid molecule is altered to reflect that of the host organism desired for expression of that nucleic acid molecule while also decreasing the number of potential transcription regulatory sequences relative to the parent nucleic acid molecule.

Thus, the invention provides a method to prepare a synthetic nucleic acid molecule comprising an open reading frame. The method comprises altering (e.g., decreasing or eliminating) a plurality of transcription regulatory sequences in a parent (wild type or a synthetic) nucleic acid sequence that encodes a polypeptide having at least 100 amino acids to yield a synthetic nucleic acid molecule which has a decreased number of transcription regulatory sequences and which preferably encodes the same amino acids as the parent nucleic acid molecule. The transcription regulatory sequences are selected from the group consisting of transcription factor binding sequences, intron splice sites, poly(A) addition sites, enhancer sequences and promoter sequences, and the resulting synthetic nucleic acid molecule has at least 3-fold fewer, preferably 5-fold fewer, transcription regulatory sequences relative to the parent nucleic acid sequence. The method also comprises altering greater than 25% of the codons in the synthetic nucleic acid sequence which has a decreased number of transcription regulatory sequences to yield a further synthetic nucleic acid molecule, wherein the codons that are altered encode the same amino acids as those in the corresponding position in the synthetic nucleic acid molecule which has a decreased number of transcription regulatory sequences and/or in the parent nucleic acid sequence. Preferably, the codons which are altered do not result in an increase in transcriptional regulatory sequences. Preferably, the further synthetic nucleic acid molecule encodes a polypeptide that has at least 85%, preferably 90%, and most preferably 95% or 99% contiguous amino acid sequence identity to the amino acid sequence of the polypeptide encoded by the parent nucleic acid sequence.

Alternatively, the method comprises altering greater than 25% of the codons in a parent nucleic acid sequence which encodes a polypeptide having at least 100 amino acids to yield a codon-altered synthetic nucleic acid molecule, wherein the codons that are altered encode the same amino acids as those present in the corresponding positions in the parent nucleic acid sequence. Then, a plurality of transcription regulatory sequences in the codon-altered synthetic nucleic acid molecule are altered to yield a further synthetic nucleic acid molecule. Preferably, the codons which are altered do not result in an increase in transcriptional regulatory sequences. Also, preferably, the further synthetic nucleic acid molecule encodes a polypeptide that has at least 85%, preferably 90%, and most preferably 95% or 99% contiguous amino acid sequence identity to the amino acid sequence of the polypeptide encoded by the parent nucleic acid sequence. Also provided is a synthetic (including a further synthetic) nucleic acid molecule prepared by the methods of the invention.

As described hereinbelow, the methods of the invention were employed with click beetle luciferase and *Renilla* luciferase nucleic acid sequences. While both of these nucleic acid molecules encode luciferase proteins, they are from entirely different families and are widely separated evolutionarily. These proteins have unrelated amino acid sequences, protein structures, and they utilize dissimilar chemical substrates. The fact that they share the name "luciferase" should not be interpreted to mean that they are from the same family, or even largely similar families. The methods produced synthetic luciferase nucleic acid molecules which exhibited significantly enhanced levels of mammalian expression without negatively effecting other desirable physical or biochemical properties (including protein half-life) and which were also largely devoid of known transcription regulatory elements.

The invention also provides at least two synthetic nucleic acid molecules that encode highly related polypeptides, but which synthetic nucleic acid molecules have an increased number of nucleotide differences relative to each other. These differences decrease the recombination frequency between the two synthetic nucleic acid molecules when those molecules are both present in a cell (i.e., they are "codon distinct" versions of a synthetic nucleic acid molecule). Thus, the invention provides a method for preparing at least two synthetic nucleic acid molecules that are codon distinct versions of a parent nucleic acid sequence that encodes a polypeptide. The method comprises altering a parent nucleic acid sequence to yield a first synthetic nucleic acid molecule having an increased number of a first plurality of codons that are employed more frequently in a selected host cell relative to the number of those codons present in the parent nucleic acid sequence. Optionally, the first synthetic nucleic acid molecule also has a decreased number of transcription regulatory sequences relative to the parent nucleic acid sequence. The parent nucleic acid sequence is also altered to yield a second synthetic nucleic acid molecule having an increased number of a second plurality of codons that are employed more frequently in the host cell relative to the number of those codons in the parent nucleic acid sequence, wherein the first plurality of codons is different than the second plurality of codons, and wherein the first and the second synthetic nucleic acid molecules preferably encode the same polypeptide. Optionally, the second synthetic nucleic acid molecule has a decreased number of transcription regulatory sequences relative to the parent nucleic acid sequence. Either or both synthetic molecules can then be further modified.

Clearly, the present invention has applications with many genes and across many fields of science including, but not limited to, life science research, agrigenetics, genetic therapy, developmental science and pharmaceutical development.

Brief Description of the Figures

Figure 1. Codons and their corresponding amino acids.

Figure 2. A nucleotide sequence comparison of a yellow-green (YG) click beetle luciferase nucleic acid sequence (YG #81-6G01; SEQ ID NO:2) and various synthetic green (GR) click beetle luciferase nucleic acid sequences (GRver1, SEQ ID NO:3; GRver2, SEQ ID NO:4; GRver3, SEQ ID NO:5; GRver4, SEQ ID NO:6; GRver5, SEQ ID NO:7; GR6, SEQ ID NO:8; GRver5.1, SEQ ID NO:9) and various red (RD) click beetle luciferase nucleic acid sequences (RDver1, SEQ ID NO:10; RDver2, SEQ ID NO:11; RDver3, SEQ ID NO:12; RDver4, SEQ ID NO:13; RDver5, SEQ ID NO:14; RD7, SEQ ID NO:15; RDver5.1, SEQ ID NO:16; RDver5.2, SEQ ID NO:17; RD156-1H9, SEQ ID NO:18). The nucleotides enclosed in boxes are nucleotides that differ from the nucleotide present at the homologous position in SEQ ID NO:2.

Figure 3. An amino acid sequence comparison of a YG click beetle luciferase amino acid sequence (YG#81-6G01, SEQ ID NO:24) and various synthetic GR click beetle luciferase amino acid sequences (GRver1, SEQ ID NO:25; GRver2, SEQ ID NO:26; GRver3, SEQ ID NO:27; GRver4, SEQ ID NO:28; GRver5, SEQ ID NO:29; GR6, SEQ ID NO:30; GRver5.1, SEQ ID NO:31) and various red (RD) click beetle luciferase amino acid sequences (RDver1, SEQ ID NO:32; RDver2, SEQ ID NO:33; RDver3, SEQ ID NO:34; RDver4, SEQ ID NO:218; RDver5, SEQ ID NO:219; RD7, SEQ ID NO:220; RDver5.1, SEQ ID

NO:221; RDver5.2, SEQ ID NO:222; RD156-1H9, SEQ ID NO:223). All amino acid sequences are inferred from the corresponding nucleotide sequence. The amino acids enclosed in boxes are amino acids that differ from the amino acid present at the homologous position in SEQ ID NO:24.

Figure 4. Codon usage in YG#81-6G01, GRver1, RDver1, GRver5, and RDver5, and humans (HUM) and relative codon usage in YG#81-6G01, GRver5, RDver5, and humans.

Figure 5. Codon usage summaries for YG#81-6G01 (Figure 5A), and GR/RD synthetic nucleic acid sequences, GRver1 (Figure 5B), RDver1 (Figure 5C), GRver2 (Figure 5D), RDver2 (Figure 5E), GRver3 (Figure 5F), RDver3 (Figure 5G), GRver4 (Figure 5H), RDver4 (Figure 5I), GRver5 (Figure 5J), RDver5 (5K).

Figure 6. Oligonucleotides employed to prepare synthetic GR/RD luciferase genes (SEQ ID Nos. 35-245).

Figure 7. A nucleotide sequence comparison of a wild type *Renilla reniformis* luciferase nucleic acid sequence Genbank Accession No. M63501 (RELLUC, SEQ ID NO:19) and various synthetic *Renilla* luciferase nucleic acid sequences (Rlucver1, SEQ ID NO:20; Rlucver2, SEQ ID NO:21; Rluc-final, SEQ ID NO:22). The nucleotides enclosed in boxes are nucleotides that differ from the nucleotide present at the homologous position in SEQ ID NO:19.

Figure 8. An amino acid sequence comparison of a wild type *Renilla reniformis* luciferase amino acid sequence (RELLUC, SEQ ID NO:224) and various synthetic *Renilla reniformis* luciferase amino acid sequences (Rlucver1, SEQ ID NO:225; Rlucver2, SEQ ID NO:226; Rluc-final, SEQ ID NO:227). All amino acid sequences are inferred from the corresponding nucleotide sequence. The amino acids enclosed in boxes are amino acids that differ from the amino acid present at the homologous position in SEQ ID NO:224.

Figure 9. Codon usage in wild-type (A) versus synthetic (B) *Renilla* luciferase genes. For codon usage in selected organisms, see, e.g., Wada et al., 1990; Sharp et al., 1988; Aota et al., 1988; and Sharp et al., 1987, and for plant codons, Murray et al. 1989.

Figure 10. Oligonucleotides employed to prepare synthetic *Renilla* luciferase gene (SEQ ID Nos. 246-292).

Figure 11. A nucleotide sequence comparison of a wild type yellow-green (YG) click beetle luciferase nucleic acid sequence (LUCPPLYG, SEQ ID NO:1) and the synthetic green click beetle luciferase nucleic acid sequences (GRver5.1, SEQ ID NO:9) and the synthetic red click beetle luciferase nucleic acid sequences (RD156-1H9, SEQ ID NO:18). The nucleotides enclosed in boxes are nucleotides that differ from the nucleotide present at the homologous position in SEQ ID NO:1. Both synthetic sequences have a codon composition that differs from LUCPPLYG at more than 25% of the codons and have at least 3-fold fewer transcription regulatory sequences relative to a random selection of codons at the codons which differ.

Figure 12. An amino acid sequence comparison of a wild type YG click beetle luciferase amino acid sequence (LUCPPLYG, SEQ ID NO:23) and the synthetic GR click beetle luciferase amino acid sequences (GRver5.1, SEQ ID NO:31) and the red (RD) click beetle luciferase amino acid sequences (RD156-1H9, SEQ ID NO:223). All amino acid sequences are inferred from the corresponding nucleotide sequence. The amino acids enclosed in boxes are amino acids that differ from the amino acid present at the homologous position in SEQ ID NO:23.

Figure 13. pRL vector series. All of the vectors contain the *Renilla* wild type or synthetic gene as further described herein. Figure 13A illustrates the *Renilla* luciferase gene in the pGL3 vectors (Promega Corp.) Figure 13B illustrates the *Renilla* luciferase co-reporter vector series. pRL-TK has the herpes simplex virus (HSV) tk promoter; pRL-SV40 has the SV40 virus early enhancer/promoter; pRL-CMV has the cytomegalovirus (CMV) enhancer and immediate early promoter; pRL-null has MCS (multiple cloning sites) but no promoter or enhancer; pRL-TK(Int $^-$) has HSV/tk promoter without an intron that is present in the other plasmids; pR-GL3B has the pGL-3 Basic backbone (Promega Corp.); pR-GL3 TK has the pGL3-Basic backbone with an HSV tk promoter.

Figure 14. Half-life of synthetic (Rluc-final) and native *Renilla* luciferases in CHO cells.

Figures 15A-B. *In vitro* transcription/translation of *Renilla* luciferase nucleic acid sequences. A) t = 0-60 minutes; B) linear range.

Figures 15C-D. *In vitro* translation of native and synthetic (Rluc-final) *Renilla* luciferase RNAs in a rabbit reticulocyte lysate. RNA was quantitated and the same amount was employed as in the translation reaction shown in Figures 15A-B. C) t = 0-60 minutes; D) linear range.

Figures 15E-F. Translation of native and synthetic (Rluc-final) *Renilla* RNAs in a wheat germ extract. E) t = 0-60 minutes; F) linear range.

Figure 16. High expression from a synthetic *Renilla* nucleic acid sequence reduces the risk of promoter interference in a co-transfection assay. CHO cells were co-transfected with a constant amount (50 ng) of firefly luciferase expression vector (pGL3 control vector, with SV40 promoter and enhancer; Luc+) and a pRL vector having a native (0 ng, 50 ng, 100 ng, 500 ng, 1 µg or 2 µg) or synthetic (0 ng, 5 ng, 10 ng, 50 ng, 100 ng or 200 ng) *Renilla* luciferase gene.

Figures 17A-B. Illustrates the reactions catalyzed by firefly and click beetle (17A), and *Renilla* (17B) luciferases.

Figure 18. Nucleotide and inferred amino acid sequence of click beetle luciferases in pGL3 vectors (GRver5.1 in pGL3, SEQ ID NO:297 encoding SEQ ID NO:298; RDver5.1 in pGL3, SEQ ID NO:299 encoding SEQ ID NO:300; and RD156-1H9 in pGL3, SEQ ID NO:301 encoding SEQ ID NO:302). To clone GRver5.1, RDver5.1, and RD156-1H9 nucleic acid sequences into pGL3 vectors, an oligonucleotide having an *Nco* I site at the initiation codon was employed, which resulted in an amino acid substitution at position 2 to valine.

Detailed Description of the Invention

Definitions

The term "gene" as used herein, refers to a DNA sequence that comprises coding sequences necessary for the production of a polypeptide or protein precursor.

The polypeptide can be encoded by a full length coding sequence or by any portion of the coding sequence, as long as the desired protein activity is retained.

A "nucleic acid", as used herein, is a covalently linked sequence of nucleotides in which the 3′ position of the pentose of one nucleotide is joined by a phosphodiester group to the 5′ position of the pentose of the next, and in which the nucleotide residues (bases) are linked in specific sequence, i.e., a linear order of nucleotides. A "polynucleotide", as used herein, is a nucleic acid containing a sequence that is greater than about 100 nucleotides in length. An "oligonucleotide", as used herein, is a short polynucleotide or a portion of a polynucleotide. An oligonucleotide typically contains a sequence of about two to about one hundred bases. The word "oligo" is sometimes used in place of the word "oligonucleotide".

Nucleic acid molecules are said to have a "5′-terminus" (5′ end) and a "3′-terminus" (3′ end) because nucleic acid phosphodiester linkages occur to the 5′ carbon and 3′ carbon of the pentose ring of the substituent mononucleotides. The end of a polynucleotide at which a new linkage would be to a 5′ carbon is its 5′ terminal nucleotide. The end of a polynucleotide at which a new linkage would be to a 3′ carbon is its 3′ terminal nucleotide. A terminal nucleotide, as used herein, is the nucleotide at the end position of the 3′- or 5′-terminus.

DNA molecules are said to have "5′ ends" and "3′ ends" because mononucleotides are reacted to make oligonucleotides in a manner such that the 5′ phosphate of one mononucleotide pentose ring is attached to the 3′ oxygen of its neighbor in one direction via a phosphodiester linkage. Therefore, an end of an oligonucleotides referred to as the "5′ end" if its 5′ phosphate is not linked to the 3′ oxygen of a mononucleotide pentose ring and as the "3′ end" if its 3′ oxygen is not linked to a 5′ phosphate of a subsequent mononucleotide pentose ring.

As used herein, a nucleic acid sequence, even if internal to a larger oligonucleotide or polynucleotide, also may be said to have 5′ and 3′ ends. In either a linear or circular DNA molecule, discrete elements are referred to as being "upstream" or 5′ of the "downstream" or 3′ elements. This terminology reflects the fact that transcription proceeds in a 5′ to 3′ fashion along the DNA strand. Typically, promoter and enhancer elements that direct transcription of a linked gene are generally located 5′ or upstream of the coding region. However, enhancer elements can exert their effect even when located 3′ of the promoter element and the coding region. Transcription termination and polyadenylation signals are located 3′ or downstream of the coding region.

The term “codon” as used herein, is a basic genetic coding unit, consisting of a sequence of three nucleotides that specify a particular amino acid to be incorporation into a polypeptide chain, or a start or stop signal. Figure 1 contains a codon table. The term "coding region" when used in reference to structural gene refers to the nucleotide sequences that encode the amino acids found in the nascent polypeptide as a result of translation of a mRNA molecule. Typically, the coding region is bounded on the 5′ side by the nucleotide triplet "ATG" which encodes the initiator methionine and on the 3′ side by a stop codon (e.g., TAA, TAG, TGA). In some cases the coding region is also known to initiate by a nucleotide triplet "TTG".

By “protein” and “polypeptide” is meant any chain of amino acids, regardless of length or post-translational modification (e.g., glycosylation or phosphorylation). The synthetic genes of the invention may also encode a variant of a naturally-occurring protein or polypeptide fragment thereof. Preferably, such a protein polypeptide has an amino acid sequence that is at least 85%, preferably 90%, and most preferably 95% or 99% identical to the amino acid sequence of the naturally-occurring (native) protein from which it is derived.

Polypeptide molecules are said to have an "amino terminus" (N-terminus) and a "carboxy terminus" (C-terminus) because peptide linkages occur between the backbone amino group of a first amino acid residue and the backbone carboxyl group of a second amino acid residue. The terms "N-terminal" and "C-terminal" in reference to polypeptide sequences refer to regions of polypeptides including portions of the N-terminal and C-terminal regions of the polypeptide, respectively. A sequence that includes a portion of the N-terminal region of polypeptide includes amino acids predominantly from the N-terminal half of the polypeptide chain, but is not limited to such sequences. For example, an N-terminal sequence may include an interior portion of the polypeptide sequence including bases from both the N-terminal and C-terminal halves of the polypeptide. The same applies to C-terminal regions. N-terminal and C-terminal regions may, but need not, include the amino acid defining the ultimate N-terminus and C-terminus of the polypeptide, respectively.

The term "wild type" as used herein, refers to a gene or gene product that has the characteristics of that gene or gene product isolated from a naturally occurring source. A wild type gene is that which is most frequently observed in a population and is thus arbitrarily designated the "wild type" form of the gene. In contrast, the term "mutant" refers to a gene or gene product that displays modifications in sequence and/or functional properties (i.e., altered characteristics) when compared to the wild type gene or gene product. It is noted that naturally-occurring mutants can be isolated; these are identified by the fact that they have altered characteristics when compared to the wild type gene or gene product.

The terms "complementary" or "complementarity" are used in reference to a sequence of nucleotides related by the base-pairing rules. For example, for the sequence 5′ "A-G-T" 3′, is complementary to the sequence 3′ "T-C-A" 5′. Complementarity may be "partial," in which only some of the nucleic acids' bases are matched according to the base pairing rules. Or, there may be "complete" or "total" complementarity between the nucleic acids. The degree of complementarity between nucleic acid strands has significant effects on the efficiency and strength of hybridization between nucleic acid strands. This is of particular importance in amplification reactions, as well as detection methods which depend upon hybridization of nucleic acids.

The term "recombinant protein" or "recombinant polypeptide" as used herein refers to a protein molecule expressed from a recombinant DNA molecule. In contrast, the term "native protein" is used herein to indicate a protein isolated from a naturally occurring (i.e., a nonrecombinant) source. Molecular biological techniques may be used to produce a recombinant form of a protein with identical properties as compared to the native form of the protein.

The terms "fusion protein" and "fusion partner" refer to a chimeric protein containing the protein of interest (e.g., luciferase) joined to an exogenous protein fragment (e.g., a fusion partner which consists of a non-luciferase protein). The fusion partner may enhance the solubility of protein as expressed in a host cell, may, for example, provide an affinity tag to allow purification of the recombinant fusion protein from the host cell or culture supernatant, or both. If desired, the fusion partner may be removed from the protein of interest by a variety of enzymatic or chemical means known to the art.

The terms "cell," "cell line," "host cell," as used herein, are used interchangeably, and all such designations include progeny or potential progeny of these designations. By “transformed cell” is meant a cell into which (or into an ancestor of which) has been introduced a DNA molecule comprising a synthetic gene. Optionally, a synthetic gene of the invention may be introduced into a suitable cell line so as to create a stably-transfected cell line capable of producing the protein or polypeptide encoded by the synthetic gene. Vectors , cells, and methods for constructing such cell lines are well known in the art, e.g. in Ausubel, et al. (infra). The words "transformants" or "transformed cells" include the primary transformed cells derived from the originally transformed cell without regard to the number of transfers. All progeny may not be precisely identical in DNA content, due to deliberate or inadvertent mutations. Nonetheless, mutant progeny that have the same functionality as screened for in the originally transformed cell are included in the definition of transformants.

Nucleic acids are known to contain different types of mutations. A "point" mutation refers to an alteration in the sequence of a nucleotide at a single base position from the wild type sequence. Mutations may also refer to insertion or deletion of one or more bases, so that the nucleic acid sequence differs from the wild-type sequence.

The term "homology" refers to a degree of complementarity. There may be partial homology or complete homology (i.e., identity). Homology is often measured using sequence analysis software (e.g., Sequence Analysis Software Package of the Genetics Computer Group. University of Wisconsin Biotechnology Center. 1710 University Avenue. Madison, WI 53705). Such software matches similar sequences by assigning degrees of homology to various substitutions, deletions, insertions, and other modifications. Conservative substitutions typically include substitutions within the following groups: glycine, alanine; valine, isoleucine, leucine; aspartic acid, glutamic acid, asparagine, glutamine; serine, threonine; lysine, arginine; and phenylalanine, tyrosine.

A “partially complementary” sequence is one that at least partially inhibits a completely complementary sequence from hybridizing to a target nucleic acid is referred to using the functional term "substantially homologous." The inhibition of hybridization of the completely complementary sequence to the target sequence may be examined using a hybridization assay (Southern or Northern blot, solution hybridization and the like) under conditions of low stringency. A substantially homologous sequence or probe will compete for and inhibit the binding (i.e., the hybridization) of a completely homologous to a target under conditions of low stringency. This is not to say that conditions of low stringency are such that non-specific binding is permitted; low stringency conditions require that the binding of two sequences to one another be a specific (i.e., selective) interaction. The absence of non-specific binding may be tested by the use of a second target which lacks even a partial degree of complementarity (e.g., less than about 30% identity). In this case, in the absence of non-specific binding, the probe will not hybridize to the second non-complementary target.

When used in reference to a double-stranded nucleic acid sequence such as a cDNA or a genomic clone, the term "substantially homologous" refers to any probe which can hybridize to either or both strands of the double-stranded nucleic acid sequence under conditions of low stringency as described herein.

"Probe" refers to an oligonucleotide designed to be sufficiently complementary to a sequence in a denatured nucleic acid to be probed (in relation to its length) to be bound under selected stringency conditions.

"Hybridization" and "binding" in the context of probes and denature melted nucleic acid are used interchangeably. Probes which are hybridized or bound to denatured nucleic acid are base paired to complementary sequences in the polynucleotide. Whether or not a particular probe remains base paired with the polynucleotide depends on the degree of complementarity, the length of the probe, and the stringency of the binding conditions. The higher the stringency, the higher must be the degree of complementarity and/or the longer the probe.

The term "hybridization" is used in reference to the pairing of complementary nucleic acid strands. Hybridization and the strength of hybridization (i.e., the strength of the association between nucleic acid strands) is impacted by many factors well known in the art including the degree of complementarity between the nucleic acids, stringency of the conditions involved affected by such conditions as the concentration of salts, the Tm (melting temperature) of the formed hybrid, the presence of other components (e.g., the presence or absence of polyethylene glycol), the molarity of the hybridizing strands and the G:C content of the nucleic acid strands.

The term "stringency" is used in reference to the conditions of temperature, ionic strength, and the presence of other compounds, under which nucleic acid hybridizations are conducted. With "high stringency" conditions, nucleic acid base pairing will occur only between nucleic acid fragments that have a high frequency of complementary base sequences. Thus, conditions of "medium" or "low" stringency are often required when it is desired that nucleic acids which are not completely complementary to one another be hybridized or annealed together. The art knows well that numerous equivalent conditions can be employed to comprise medium or low stringency conditions. The choice of hybridization conditions is generally evident to one skilled in the art and is usually guided by the purpose of the hybridization, the type of hybridization (DNA-DNA or DNA-RNA), and the level of desired relatedness between the sequences (e.g., Sambrook et al., 1989; Nucleic Acid Hybridization, A Practical Approach, IRL Press, Washington D.C., 1985, for a general discussion of the methods).

The stability of nucleic acid duplexes is known to decrease with an increased number of mismatched bases, and further to be decreased to a greater or lesser degree depending on the relative positions of mismatches in the hybrid duplexes. Thus, the stringency of hybridization can be used to maximize or minimize stability of such duplexes. Hybridization stringency can be altered by: adjusting the temperature of hybridization; adjusting the percentage of helix destabilizing agents, such as formamide, in the hybridization mix; and adjusting the temperature and/or salt concentration of the wash solutions. For filter hybridizations, the final stringency of hybridizations often is determined by the salt concentration and/or temperature used for the post-hybridization washes.

"High stringency conditions" when used in reference to nucleic acid hybridization comprise conditions equivalent to binding or hybridization at 42°C in a solution consisting of 5X SSPE (43.8 g/l NaCl, 6.9 g/l NaH_2PO_4 H_2O and 1.85 g/l EDTA, pH adjusted to 7.4 with NaOH), 0.5% SDS, 5X Denhardt's reagent and 100 μg/ml denatured salmon sperm DNA followed by washing in a solution comprising 0.1X SSPE, 1.0% SDS at 42°C when a probe of about 500 nucleotides in length is employed.

"Medium stringency conditions" when used in reference to nucleic acid hybridization comprise conditions equivalent to binding or hybridization at 42°C in a solution consisting of 5X SSPE (43.8 g/l NaCl, 6.9 g/l NaH_2PO_4 H_2O and 1.85 g/l EDTA, pH adjusted to 7.4 with NaOH), 0.5% SDS, 5X Denhardt's reagent and 100 μg/ml denatured salmon sperm DNA followed by washing in a solution comprising 1.0X SSPE, 1.0% SDS at 42°C when a probe of about 500 nucleotides in length is employed.

"Low stringency conditions" comprise conditions equivalent to binding or hybridization at 42°C in a solution consisting of 5X SSPE (43.8 g/l NaCl, 6.9 g/l NaH_2PO_4 H_2O and 1.85 g/l EDTA, pH adjusted to 7.4 with NaOH), 0.1% SDS, 5X

Denhardt's reagent [50X Denhardt's contains per 500 ml: 5 g Ficoll (Type 400, Pharmacia), 5 g BSA (Fraction V; Sigma)] and 100 g/ml denatured salmon sperm DNA followed by washing in a solution comprising 5X SSPE, 0.1% SDS at 42°C when a probe of about 500 nucleotides in length is employed.

The term "T_m" is used in reference to the "melting temperature". The melting temperature is the temperature at which 50% of a population of double-stranded nucleic acid molecules becomes dissociated into single strands. The equation for calculating the T_m of nucleic acids is well-known in the art. The Tm of a hybrid nucleic acid is often estimated using a formula adopted from hybridization assays in 1 M salt, and commonly used for calculating Tm for PCR primers: [(number of A + T) x 2°C + (number of G+C) x 4°C]. (C.R. Newton et al., PCR, 2nd Ed., Springer-Verlag (New York, 1997), p. 24). This formula was found to be inaccurate for primers longer than 20 nucleotides. (Id.) Another simple estimate of the T_m value may be calculated by the equation: $T_m = 81.5 + 0.41(\% G + C)$, when a nucleic acid is in aqueous solution at 1 M NaCl. (e.g., Anderson and Young, Quantitative Filter Hybridization, in Nucleic Acid Hybridization, 1985). Other more sophisticated computations exist in the art which take structural as well as sequence characteristics into account for the calculation of T_m. A calculated T_m is merely an estimate; the optimum temperature is commonly determined empirically.

The term "isolated" when used in relation to a nucleic acid, as in "isolated oligonucleotide" or "isolated polynucleotide" refers to a nucleic acid sequence that is identified and separated from at least one contaminant with which it is ordinarily associated in its source. Thus, an isolated nucleic acid is present in a form or setting that is different from that in which it is found in nature. In contrast, non-isolated nucleic acids (e.g., DNA and RNA) are found in the state they exist in nature. For example, a given DNA sequence (e.g., a gene) is found on the host cell chromosome in proximity to neighboring genes; RNA sequences (e.g., a specific mRNA sequence encoding a specific protein), are found in the cell as a mixture with numerous other mRNAs that encode a multitude of proteins. However, isolated nucleic acid includes, by way of example, such nucleic acid in cells ordinarily expressing that nucleic acid where the nucleic acid is in a chromosomal location different from that of natural cells, or is otherwise flanked by a different nucleic acid sequence than that found in nature. The isolated nucleic acid or oligonucleotide may be present in single-stranded or double-stranded form. When an isolated nucleic acid or oligonucleotide is to be utilized to express a protein, the oligonucleotide contains at a minimum, the sense or coding strand (i.e., the oligonucleotide may single-stranded), but may contain both the sense and anti-sense strands (i.e., the oligonucleotide may be double-stranded).

The term "isolated" when used in relation to a polypeptide, as in "isolated protein" or "isolated polypeptide" refers to a polypeptide that is identified and separated from at least one contaminant with which it is ordinarily associated in its source. Thus, an isolated polypeptide is present in a form or setting that is different from that in which it is found in nature. In contrast, non-isolated polypeptides (e.g., proteins and enzymes) are found in the state they exist in nature.

The term "purified" or "to purify" means the result of any process that removes some of a contaminant from the component of interest, such as a protein or nucleic acid. The percent of a purified component is thereby increased in the sample.

The term "operably linked" as used herein refer to the linkage of nucleic acid sequences in such a manner that a nucleic acid molecule capable of directing the transcription of a given gene and/or the synthesis of a desired protein molecule is produced. The term also refers to the linkage of sequences encoding amino acids in such a manner that a functional (e.g., enzymatically active, capable of binding to a binding partner, capable of inhibiting, etc.) protein or polypeptide is produced.

The term "recombinant DNA molecule" means a hybrid DNA sequence comprising at least two nucleotide sequences not normally found together in nature.

The term "vector" is used in reference to nucleic acid molecules into which fragments of DNA may be inserted or cloned and can be used to transfer DNA segment(s) into a cell and capable of replication in a cell. Vectors may be derived from plasmids, bacteriophages, viruses, cosmids, and the like.

The terms "recombinant vector" and "expression vector" as used herein refer to DNA or RNA sequences containing a desired coding sequence and appropriate DNA or RNA sequences necessary for the expression of the operably linked coding sequence in a particular host organism. Prokaryotic expression vectors include a promoter, a ribosome binding site, an origin of replication for autonomous replication in a host cell and possibly other sequences, e.g. an optional operator sequence, optional restriction enzyme sites. A promoter is defined as a DNA sequence that directs RNA polymerase to bind to DNA and to initiate RNA synthesis. Eukaryotic expression vectors include a promoter, optionally a polyadenlyation signal and optionally an enhancer sequence.

The term "a polynucleotide having a nucleotide sequence encoding a gene," means a nucleic acid sequence comprising the coding region of a gene, or in other words the nucleic acid sequence which encodes a gene product. The coding region may be present in either a cDNA, genomic DNA or RNA form. When present in a DNA form, the oligonucleotide may be single-stranded (i.e., the sense strand) or double-stranded. Suitable control elements such as enhancers/promoters, splice junctions, polyadenylation signals, etc. may be placed in close proximity to the coding region of the gene if needed to permit proper initiation of transcription and/or correct processing of the primary RNA transcript. Alternatively, the coding region utilized in the expression vectors of the present invention may contain endogenous enhancers/promoters, splice junctions, intervening sequences, polyadenylation signals, etc. In further embodiments, the coding region may contain a combination of both endogenous and exogenous control elements.

The term "transcription regulatory element" or “transcription regulatory sequence” refers to a genetic element or sequence that controls some aspect of the expression of nucleic acid sequence(s). For example, a promoter is a regulatory element that facilitates the initiation of transcription of an operably linked coding region. Other regulatory elements include, but are not limited to, transcription factor binding sites, splicing signals, polyadenylation signals, termination signals and enhancer elements.

Transcriptional control signals in eukaryotes comprise "promoter" and "enhancer" elements. Promoters and enhancers consist of short arrays of DNA sequences that interact specifically with cellular proteins involved in transcription (Maniatis et al., 1987). Promoter and enhancer elements have been isolated from a variety of eukaryotic sources including genes in yeast, insect and mammalian cells. Promoter and enhancer elements have also been isolated from viruses and analogous control elements, such as promoters, are also found in prokaryotes. The selection of a particular promoter and enhancer depends on the cell type used to express the protein of interest. Some eukaryotic promoters and enhancers have a broad host range while others are functional in a limited subset of cell types (for review, see Voss et al., 1986; and Maniatis et al., 1987. For example, the SV40 early gene enhancer is very active in a wide variety of cell types from many mammalian species and has been widely used for the expression of proteins in mammalian cells (Dijkema et al., 1985). Two other examples of promoter/enhancer elements active in a broad range of mammalian cell types are those from the human elongation factor 1 gene (Uetsuki et al., 1989; Kim, et al., 1990; and Mizushima and Nagata, 1990) and the long terminal repeats of the Rous sarcoma virus (Gorman et al., 1982); and the human cytomegalovirus (Boshart et al., 1985).

The term "promoter/enhancer" denotes a segment of DNA containing sequences capable of providing both promoter and enhancer functions (i.e., the functions provided by a promoter element and an enhancer element as described above). For example, the long terminal repeats of retroviruses contain both promoter and enhancer functions. The enhancer/promoter may be "endogenous" or "exogenous" or "heterologous." An "endogenous" enhancer/promoter is one that is naturally linked with a given gene in the genome. An "exogenous" or "heterologous" enhancer/promoter is one that is placed in juxtaposition to a gene by means of genetic manipulation (i.e., molecular biological techniques) such that transcription of the gene is directed by the linked enhancer/promoter.

The presence of "splicing signals" on an expression vector often results in higher levels of expression of the recombinant transcript in eukaryotic host cells.

Splicing signals mediate the removal of introns from the primary RNA transcript and consist of a splice donor and acceptor site (Sambrook, et al., Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor Laboratory Press, New York , 1989, pp. 16.7-16.8). A commonly used splice donor and acceptor site is the splice junction from the 16S RNA of SV40.

Efficient expression of recombinant DNA sequences in eukaryotic cells requires expression of signals directing the efficient termination and polyadenylation of the resulting transcript. Transcription termination signals are generally found downstream of the polyadenylation signal and are a few hundred nucleotides in length. The term "poly(A) site" or "poly(A) sequence" as used herein denotes a DNA sequence which directs both the termination and polyadenylation of the nascent RNA transcript. Efficient polyadenylation of the recombinant transcript is desirable, as transcripts lacking a poly(A) tail are unstable and are rapidly degraded. The poly(A) signal utilized in an expression vector may be "heterologous" or "endogenous." An endogenous poly(A) signal is one that is found naturally at the 3′ end of the coding region of a given gene in the genome. A heterologous poly(A) signal is one which has been isolated from one gene and positioned 3′ to another gene. A commonly used heterologous poly(A) signal is the SV40 poly(A) signal. The SV40 poly(A) signal is contained on a 237 bp *Bam*H I/*Bcl* I restriction fragment and directs both termination and polyadenylation (Sambrook, supra, at 16.6-16.7).

Eukaryotic expression vectors may also contain "viral replicons "or "viral origins of replication." Viral replicons are viral DNA sequences which allow for the extrachromosomal replication of a vector in a host cell expressing the appropriate replication factors. Vectors containing either the SV40 or polyoma virus origin of replication replicate to high copy number (up to 10^4 copies/cell) in cells that express the appropriate viral T antigen. In contrast, vectors containing the replicons from bovine papillomavirus or Epstein-Barr virus replicate extrachromosomally at low copy number (about 100 copies/cell).

The term "*in vitro*" refers to an artificial environment and to processes or reactions that occur within an artificial environment. *In vitro* environments include, but are not limited to, test tubes and cell lysates. The term "*in situ*" refers to cell culture. The term "*in vivo*" refers to the natural environment (e.g., an animal or a cell) and to processes or reaction that occur within a natural environment.

The term "expression system" refers to any assay or system for determining (e.g., detecting) the expression of a gene of interest. Those skilled in the field of molecular biology will understand that any of a wide variety of expression systems may be used. A wide range of suitable mammalian cells are available from a wide range of source (e.g., the American Type Culture Collection, Rockland, MD). The method of transformation or transfection and the choice of expression vehicle will depend on the host system selected. Transformation and transfection methods are described, e.g., in Ausubel, et al., Current Protocols in Molecular Biology. John Wiley & Sons, New York. 1992. Expression systems include *in vitro* gene expression assays where a gene of interest (e.g., a reporter gene) is linked to a regulatory sequence and the expression of the gene is monitored following treatment with an agent that inhibits or induces expression of the gene. Detection of gene expression can be through any suitable means including, but not limited to, detection of expressed mRNA or protein (e.g., a detectable product of a reporter gene) or through a detectable change in the phenotype of a cell expressing the gene of interest. Expression systems may also comprise assays where a cleavage event or other nucleic acid or cellular change is detected.

The term "enzyme" refers to molecules or molecule aggregates that are responsible for catalyzing chemical and biological reactions. Such molecules are typically proteins, but can also comprise short peptides, RNAs, ribozymes, antibodies, and other molecules. A molecule that catalyzes chemical and biological reactions is referred to as "having enzyme activity" or "having catalytic activity."

All amino acid residues identified herein are in the natural L-configuration. In keeping with standard polypeptide nomenclature (see J. Biol. Chem., 243, 3557 (1969)), abbreviations for amino acid residues are as shown in the following Table of Correspondence.

TABLE OF CORRESPONDENCE

1-Letter	3-Letter	AMINO ACID
Y	Tyr	L-tyrosine
G	Gly	glycine
F	Phe	L-phenylalanine
M	Met	L-methionine
A	Ala	L-alanine
S	Ser	L-serine
I	Ile	L-isoleucine
L	Leu	L-leucine
T	Thr	L-threonine
V	Val	L-valine
P	Pro	L-proline
K	Lys	L-lysine
H	His	L-histidine
Q	Gln	L-glutamine
E	Glu	L-glutamic acid
W	Trp	L-tryptophan
R	Arg	L-arginine
D	Asp	L-aspartic acid
N	Asn	L-asparagine
C	Cys	L-cysteine

The term “sequence homology” means the proportion of base matches between two nucleic acid sequences or the proportion of amino acid matches between two amino acid sequences. When sequence homology is expressed as a percentage, e.g., 50%, the percentage denotes the proportion of matches over the length of sequence from one sequence that is compared to some other sequence. Gaps (in either of the two sequences) are permitted to maximize matching; gap lengths of 15 bases or less are usually used, 6 bases or less are preferred with 2 bases or less more preferred. When using oligonucleotides as probes or treatments, the sequence homology between the target nucleic acid and the oligonucleotide sequence is generally not less than 17 target base matches out of 20 possible oligonucleotide base pair matches (85%); preferably not less than 9 matches out of 10 possible base pair matches (90%), and more preferably not less than 19 matches out of 20 possible base pair matches (95%).

Two amino acid sequences are homologous if there is a partial or complete identity between their sequences. For example, 85% homology means that 85% of the amino acids are identical when the two sequences are aligned for maximum matching. Gaps (in either of the two sequences being matched) are allowed in maximizing matching; gap lengths of 5 or less are preferred with 2 or less being more preferred. Alternatively and preferably, two protein sequences (or polypeptide sequences derived from them of at least 100 amino acids in length) are homologous, as this term is used herein, if they have an alignment score of at more than 5 (in standard deviation units) using the program ALIGN with the mutation data matrix and a gap penalty of 6 or greater. See Dayhoff, M. O., in Atlas of Protein Sequence and Structure, 1972, volume 5, National Biomedical Research Foundation, pp. 101-110, and Supplement 2 to this volume, pp. 1-10. The two sequences or parts thereof are more preferably homologous if their amino acids are greater than or equal to 85% identical when optimally aligned using the ALIGN program.

The following terms are used to describe the sequence relationships between two or more polynucleotides: “reference sequence”, “comparison window”, “sequence identity”, “percentage of sequence identity”, and “substantial identity”. A “reference sequence” is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA or gene sequence given in a sequence listing, or may comprise a complete cDNA or gene sequence. Generally, a reference sequence is at least 20 nucleotides in length, frequently at least 25 nucleotides in length, and often at least 50 nucleotides in length. Since two polynucleotides may each (1) comprise a sequence (i.e., a portion of the complete polynucleotide sequence) that is similar between the two polynucleotides, and (2) may further comprise a sequence that is divergent between the two polynucleotides, sequence comparisons between two (or more) polynucleotides are typically performed by comparing sequences of the two polynucleotides over a "comparison window" to identify and compare local regions of sequence similarity.

A "comparison window", as used herein, refers to a conceptual segment of at least 20 contiguous nucleotides and wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (i.e., gaps) of 20 percent or less as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences.

Methods of alignment of sequences for comparison are well known in the art. Thus, the determination of percent identity between any two sequences can be accomplished using a mathematical algorithm. Preferred, non-limiting examples of such mathematical algorithms are the algorithm of Myers and Miller (1988); the local homology algorithm of Smith and Waterman (1981); the homology alignment algorithm of Needleman and Wunsch (1970); the search-for-similarity-method of Pearson and Lipman (1988); the algorithm of Karlin and Altschul (1990), modified as in Karlin and Altschul (1993).

Computer implementations of these mathematical algorithms can be utilized for comparison of sequences to determine sequence identity. Such implementations include, but are not limited to: CLUSTAL in the PC/Gene program (available from Intelligenetics, Mountain View, California); the ALIGN program (Version 2.0) and GAP, BESTFIT, BLAST, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Version 8 (available from Genetics Computer Group (GCG), 575 Science Drive, Madison, Wisconsin, USA). Alignments using these programs can be performed using the default parameters. The CLUSTAL program is well described by Higgins et al. (1988); Higgins et al. (1989); Corpet et al. (1988); Huang et al. (1992); and Pearson et al. (1994). The ALIGN program is based on the algorithm of Myers and Miller, *supra*. The BLAST programs of Altschul et al. (1990), are based on the algorithm of Karlin and Altschul *supra*. To obtain gapped alignments for comparison purposes, Gapped BLAST (in BLAST 2.0) can be utilized as described in Altschul et al. (1997). Alternatively, PSI-BLAST (in BLAST 2.0) can be used to perform an iterated search that detects distant relationships between molecules. See Altschul et al., *supra*. When utilizing BLAST, Gapped BLAST, PSI-BLAST, the default parameters of the respective programs (e.g. BLASTN for nucleotide sequences, BLASTX for proteins) can be used. See http://www.ncbi.n1m.nih.gov. Alignment may also be performed manually by inspection

The term "sequence identity" means that two polynucleotide sequences are identical (i.e., on a nucleotide-by-nucleotide basis) over the window of comparison. The term "percentage of sequence identity" means that two polynucleotide sequences are identical (i.e., on a nucleotide-by-nucleotide basis) for the stated proportion of nucleotides over the window of comparison. The term "percentage of sequence identity" is calculated by comparing two optimally aligned sequences over the window of comparison, determining the number of positions at which the identical nucleic acid base (e.g., A, T, C, G, U, or I) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison (i.e., the window size), and multiplying the result by 100 to yield the percentage of sequence identity. The terms "substantial identity" as used herein denote a characteristic of a polynucleotide sequence, wherein the polynucleotide comprises a sequence that has at least 60%, preferably at least 65%, more preferably at least 70%, up to about 85%, and even more preferably at least 90 to 95%, more usually at least 99%, sequence identity as compared to a reference sequence over a comparison window of at least 20 nucleotide positions, frequently over a window of at least 20-50 nucleotides, and preferably at least 300 nucleotides, wherein the percentage of sequence identity is calculated by comparing the reference sequence to the polynucleotide sequence which may include deletions or additions which total 20 percent or less of the reference sequence over the window of comparison. The reference sequence may be a subset of a larger sequence.

As applied to polypeptides, the term "substantial identity" means that two peptide sequences, when optimally aligned, such as by the programs GAP or BESTFIT using default gap weights, share at least about 85% sequence identity, preferably at least about 90% sequence identity, more preferably at least about 95 % sequence identity, and most preferably at least about 99 % sequence identity.

The Synthetic Nucleic Acid Molecules and Methods of the Invention

The invention provides compositions comprising synthetic nucleic acid molecules, as well as methods for preparing those molecules which yield synthetic nucleic acid molecules that are efficiently expressed as a polypeptide or protein with desirable characteristics including reduced inappropriate or unintended transcription characteristics when expressed in a particular cell type.

Natural selection is the hypothesis that genotype-environment interactions occurring at the phenotypic level lead to differential reproductive success of individuals and hence to modification of the gene pool of a population. It is generally accepted that the amino acid sequence of a protein found in nature has undergone optimization by natural selection. However, amino acids exist within the sequence of a protein that do not contribute significantly to the activity of the protein and these amino acids can be changed to other amino acids with little or no consequence. Furthermore, a protein may be useful outside its natural environment or for purposes that differ from the conditions of its natural selection. In these circumstances, the amino acid sequence can be synthetically altered to better adapt the protein for its utility in various applications.

Likewise, the nucleic acid sequence that encodes a protein is also optimized by natural selection. The relationship between coding DNA and its transcribed RNA is such that any change to the DNA affects the resulting RNA. Thus, natural selection works on both molecules simultaneously. However, this relationship does not exist between nucleic acids and proteins. Because multiple codons encode the same amino acid, many different nucleotide sequences can encode an identical protein. A specific protein composed of 500 amino acids can theoretically be encoded by more than 10^{150} different nucleic acid sequences.

Natural selection acts on nucleic acids to achieve proper encoding of the corresponding protein. Presumably, other properties of nucleic acid molecules are also acted upon by natural selection. These properties include codon usage frequency, RNA secondary structure, the efficiency of intron splicing, and interactions with transcription factors or other nucleic acid binding proteins. These other properties may alter the efficiency of protein translation and the resulting phenotype. Because of the redundant nature of the genetic code, these other attributes can be optimized by natural selection without altering the corresponding amino acid sequence.

Under some conditions, it is useful to synthetically alter the natural nucleotide sequence encoding a protein to better adapt the protein for alternative applications. A common example is to alter the codon usage frequency of a gene when it is expressed in a foreign host. Although redundancy in the genetic code allows amino acids to be encoded by multiple codons, different organisms favor some codons over others. The codon usage frequencies tend to differ most for organisms with widely separated evolutionary histories. It has been found that when transferring genes between evolutionarily distant organisms, the efficiency of protein translation can be substantially increased by adjusting the codon usage frequency (see U.S. Patent Nos. 5,096,825, 5,670,356 and 5,874,304).

Because of the need for evolutionary distance, the codon usage of reporter genes often does not correspond to the optimal codon usage of the experimental cells. Examples include β-galactosidase (β-*gal*) and chloramphenicol acetyltransferase (*cat*) reporter genes that are derived from *E. coli* and are commonly used in mammalian cells; the β-glucuronidase (*gus*) reporter gene that is derived from *E. coli* and commonly used in plant cells; the firefly luciferase (*luc*) reporter gene that is derived from an insect and commonly used in plant and mammalian cells; and the *Renilla* luciferase, and green fluorescent protein (*gfp*) reporter genes which are derived from coelenterates and are commonly used in plant and mammalian cells. To achieve sensitive quantitation of reporter gene expression, the activity of the gene product must not be endogenous to the experimental host cells. Thus, reporter genes are usually selected from organisms having unique and distinctive phenotypes. Consequently, these organisms often have widely separated evolutionary histories from the experimental host cells.

Previously, to create genes having a more optimal codon usage frequency but still encoding the same gene product, a synthetic nucleic acid sequence was made by replacing existing codons with codons that were generally more favorable to the experimental host cell (see U.S. Patent Nos. 5,096,825, 5,670,356 and 5,874,304.) The result was a net improvement in codon usage frequency of the synthetic gene. However, the optimization of other attributes was not considered and so these synthetic genes likely did not reflect genes optimized by natural selection.

In particular, improvements in codon usage frequency are intended only for optimization of a RNA sequence based on its role in translation into a protein. Thus, previously described methods did not address how the sequence of a synthetic gene affects the role of DNA in transcription into RNA. Most notably, consideration had not been given as to how transcription factors may interact with the synthetic DNA and consequently modulate or otherwise influence gene transcription. For genes found in nature, the DNA would be optimally transcribed by the native host cell and would yield an RNA that encodes a properly folded gene product. In contrast, synthetic genes have previously not been optimized for transcriptional characteristics. Rather, this property has been ignored or left to chance.

This concern is important for all genes, but particularly important for reporter genes, which are most commonly used to quantitate transcriptional behavior in the experimental host cells. Hundreds of transcription factors have been identified in different cell types under different physiological conditions, and likely more exist but have not yet been identified. All of these transcription factors can influence the transcription of an introduced gene. A useful synthetic reporter gene of the invention has a minimal risk of influencing or perturbing intrinsic transcriptional characteristics of the host cell because the structure of that gene has been altered. A particularly useful synthetic reporter gene will have desirable characteristics under a new set and/or a wide variety of experimental conditions. To best achieve these characteristics, the structure of the synthetic gene should have minimal potential for interacting with transcription factors within a broad range of host cells and physiological conditions. Minimizing potential interactions between a reporter gene and a host cell's endogenous transcription factors increases the value of a reporter gene by reducing the risk of inappropriate transcriptional characteristics of the gene within a particular experiment, increasing applicability of the gene in various environments, and increasing the acceptance of the resulting experimental data.

In contrast, a reporter gene comprising a native nucleotide sequence, based on a genomic or cDNA clone from the original host organism, may interact with transcription factors when expressed in an exogenous host. This risk stems from two circumstances. First, the native nucleotide sequence contains sequences that were optimized through natural selection to influence gene transcription within the native host organism. However, these sequences might also influence transcription when the gene is expressed in exogenous hosts, i.e., out of context, thus interfering with its performance as a reporter gene. Second, the nucleotide sequence may inadvertently interact with transcription factors that were not present in the native host organism, and thus did not participate in its natural selection. The probability of such inadvertent interactions increases with greater evolutionary separation between the experimental cells and the native organism of the reporter gene.

These potential interactions with transcription factors would likely be disrupted when using a synthetic reporter gene having alterations in codon usage frequency. However, a synthetic reporter gene sequence, designed by choosing codons based only on codon usage frequency, is likely to contain other unintended transcription factor binding sites since the synthetic gene has not been subjected to the benefit of natural selection to correct inappropriate transcriptional activities.

Inadvertent interactions with transcription factors could also occur whenever the encoded amino acid sequence is artificially altered, e.g., to introduce amino acid substitutions. Similarly, these changes have not been subjected to natural selection, and thus may exhibit undesired characteristics.

Thus, the invention provides a method for preparing synthetic nucleic acid sequences that reduce the risk of undesirable interactions of the nucleic acid with transcription factors when expressed in a particular host cell, thereby reducing inappropriate or unintended transcriptional characteristics. Preferably, the method yields synthetic genes containing improved codon usage frequencies for a particular host cell and with a reduced occurrence of transcription factor binding sites. The invention also provides a method of preparing synthetic genes containing improved codon usage frequencies with a reduced occurrence of transcription factor binding sites and additional beneficial structural attributes. Such additional attributes include the absence of inappropriate RNA splicing junctions, poly(A) addition signals, undesirable restriction sites, ribosomal binding sites, and secondary structural motifs such as hairpin loops.

Also provided is a method for preparing two synthetic genes encoding the same or highly similar proteins ("codon distinct" versions). Preferably, the two synthetic genes have a reduced ability to hybridize to a common polynucleotide probe sequence, or have a reduced risk of recombining when present together in living cells. To detect recombination, PCR amplification of the reporter sequences using primers complementary to flanking sequences and sequencing of the amplified sequences may be employed.

To select codons for the synthetic nucleic acid molecules of the invention, preferred codons have a relatively high codon usage frequency in a selected host cell, and their introduction results in the introduction of relatively few transcription factor binding sites, relatively few other undesirable structural attributes, and optionally a characteristic that distinguishes the synthetic gene from another gene encoding a highly similar protein. Thus, the synthetic nucleic acid product obtained by the method of the invention is a synthetic gene with improved level of expression due to improved codon usage frequency, a reduced risk of inappropriate transcriptional behavior due to a reduced number of undesirable transcription regulatory sequences, and optionally any additional characteristic due to other criteria that may be employed to select the synthetic sequence.

The invention may be employed with any nucleic acid sequence, e.g., a native sequence such as a cDNA or one which has been manipulated *in vitro*, e.g., to introduce specific alterations such as the introduction or removal of a restriction enzyme recognition site, the alteration of a codon to encode a different amino acid or to encode a fusion protein, or to alter GC or AT content (% of composition) of nucleic acid molecules. Moreover, the method of the invention is useful with any gene, but particularly useful for reporter genes as well as other genes associated with the expression of reporter genes, such as selectable markers. Preferred genes include, but are not limited to, those encoding lactamase (β-gal), neomycin resistance (Neo), CAT, GUS, galactopyranoside, GFP, xylosidase, thymidine kinase, arabinosidase and the like. As used herein, a "marker gene" or "reporter gene" is a gene that imparts a distinct phenotype to cells expressing the gene and thus permits cells having the gene to be distinguished from cells that do not have the gene. Such genes may encode either a selectable or screenable marker, depending on whether the marker confers a trait which one can 'select' for by chemical means, i.e., through the use of a selective agent (e.g., a herbicide, antibiotic, or the like), or whether it is simply a "reporter" trait that one can identify through observation or testing, i.e., by 'screening'. Elements of the present disclosure are exemplified in detail through the use of particular marker genes. Of course, many examples of suitable marker genes or reporter genes are known to the art and can be employed in the practice of the invention. Therefore, it will be understood that the following discussion is exemplary rather than exhaustive. In light of the techniques disclosed herein and the general recombinant techniques which are known in the art, the present invention renders possible the alteration of any gene.

Exemplary marker genes include, but are not limited to, a *neo* gene, a β-gal gene, a *gus* gene, a *cat* gene, a *gpt* gene, a *hyg* gene, a *hisD* gene, a *ble* gene, a *mprt* gene, a *bar* gene, a nitrilase gene, a mutant acetolactate synthase gene (ALS) or acetoacid synthase gene (AAS), a methotrexate-resistant *dhfr* gene, a dalapon dehalogenase gene, a mutated anthranilate synthase gene that confers resistance to 5-methyl tryptophan (WO 97/26366), an R-locus gene, a β-lactamase gene, a *xyl*E gene, an α-amylase gene, a tyrosinase gene, a luciferase (*luc*) gene, (e.g., a *Renilla reniformis* luciferase gene, a firefly luciferase gene, or a click beetle luciferase (*Pyrophorus plagiophthalamus*) gene), an aequorin gene, or a green fluorescent protein gene. Included within the terms selectable or screenable marker genes are also genes which encode a "secretable marker" whose secretion can be detected as a means of identifying or selecting for transformed cells. Examples include markers which encode a secretable antigen that can be identified by antibody interaction, or even secretable enzymes which can be detected by their catalytic activity. Secretable proteins fall into a number of classes, including small, diffusible proteins detectable, e.g., by ELISA, and proteins that are inserted or trapped in the cell membrane.

The method of the invention can be performed by, although it is not limited to, a recursive process. The process includes assigning preferred codons to each amino acid in a target molecule, e.g., a native nucleotide sequence, based on codon usage in a particular species, identifying potential transcription regulatory sequences such as transcription factor binding sites in the nucleic acid sequence having preferred codons, e.g., using a database of such binding sites, optionally identifying other undesirable sequences, and substituting an alternative codon (i.e., encoding the same amino acid) at positions where undesirable transcription factor binding sites or other sequences occur. For codon distinct versions, alternative preferred codons are substituted in each version. If necessary, the identification and elimination of potential transcription factor or other undesirable sequences can be repeated until a nucleotide sequence is achieved containing a maximum number of preferred codons and a minimum number of undesired sequences including transcription regulatory sequences or other undesirable sequences. Also, optionally, desired sequences, e.g., restriction enzyme recognition sites, can be introduced. After a synthetic nucleic acid molecule is designed and constructed, its properties relative to the parent nucleic acid sequence can be determined by methods well known to the art. For example, the expression of the synthetic and target nucleic acid molecules in a series of vectors in a particular cell can be compared.

Thus, generally, the method of the invention comprises identifying a target nucleic acid sequence, such as a vector backbone, a reporter gene or a selectable marker gene, and a host cell of interest, for example, a plant (dicot or monocot), fungus, yeast or mammalian cell. Preferred host cells are mammalian host cells such as CHO, COS, 293, Hela, CV-1 and NIH3T3 cells. Based on preferred codon usage in the host cell(s) and, optionally, low codon usage in the host cell(s), e.g., high usage mammalian codons and low usage *E. coli* and mammalian codons, codons to be replaced are determined. For codon distinct versions of two synthetic nucleic acid molecules, alternative preferred codons are introduced to each version. Thus, for amino acids having more than two codons, one preferred codon is introduced to one version and another preferred codon is introduced to the other version. For amino acids having six codons, the two codons with the largest number of mismatched bases are identified and one is introduced to one version and the other codon is introduced to the other version. Concurrent, subsequent or prior to selecting codons to be replaced, desired and undesired sequences, such as undesired transcriptional regulatory sequences, in the target sequence are identified. These sequences can be identified using databases and software such as EPD, NNPD, REBASE, TRANSFAC, TESS, GenePro, MAR (www.ncgr.org/MAR-search) and BCM Gene Finder, further described herein. After the sequences are identified, the modification(s) are introduced. Once a desired synthetic nucleic acid sequence is obtained, it can be prepared by methods well known to the art (such as PCR with overlapping primers), and its structural and functional properties compared to the target nucleic acid sequence, including, but not limited to, percent homology, presence or absence of certain sequences, for example, restriction sites, percent of codons changed (such as an increased or decreased usage of certain codons) and expression rates.

As described below, the method was used to create synthetic reporter genes encoding *Renilla reniformis* luciferase, and two click beetle luciferases (one emitting green light and the other emitting red light). For both systems, the synthetic genes support much greater levels of expression than the corresponding native or parent genes for the protein. In addition, the native and parent genes demonstrated anomalous transcription characteristics when expressed in mammalian cells, which were not evident in the synthetic genes. In particular, basal expression of the native or parent genes is relatively high. Furthermore, the expression is induced to very high levels by an enhancer sequence in the absence of known promoters. The synthetic genes show lower basal expression and do not show the anomalous enhancer behavior. Presumably, the enhancer is activating transcriptional elements found in the native genes that are absent in the synthetic genes. The results clearly show that the synthetic nucleic acid sequences exhibit superior performance as reporter genes.

Exemplary Uses of the Molecules of the Invention

The synthetic genes of the invention preferably encode the same proteins as their native counterpart (or nearly so), but have improved codon usage while being largely devoid of known transcription regulatory elements in the coding region. (It is recognized that a small number of amino acid changes may be desired to enhance a property of the native counterpart protein, e.g. to enhance luminescence of a luciferase.) This increases the level of expression of the protein the synthetic gene encodes and reduces the risk of anomalous expression of the protein. For example, studies of many important events of gene regulation, which may be mediated by weak promoters, are limited by insufficient reporter signals from inadequate expression of the reporter proteins. The synthetic luciferase genes described herein permit detection of weak promoter activity because of the large increase in level of expression, which enables increased detection sensitivity. Also, the use of some selectable markers may be limited by the expression of that marker in an exogenous cell. Thus, synthetic selectable marker genes which have improved codon usage for that cell, and have a decrease in other undesirable sequences, (e.g., transcription factor binding sites), can permit the use of those markers in cells that otherwise were undesirable as hosts for those markers.

Promoter crosstalk is another concern when a co-reporter gene is used to normalize transfection efficiencies. With the enhanced expression of synthetic genes, the amount of DNA containing strong promoters can be reduced, or DNA containing weaker promoters can be employed, to drive the expression of the co-reporter. In addition, there may be a reduction in the background expression from the synthetic reporter genes of the invention. This characteristic makes synthetic reporter genes more desirable by minimizing the sporadic expression from the genes and reducing the interference resulting from other regulatory pathways.

The use of reporter genes in imaging systems, which can be used for *in vivo* biological studies or drug screening, is another use for the synthetic genes of the invention. Due to their increased level of expression, the protein encoded by a synthetic gene is more readily detectable by an imaging system. In fact, using a synthetic *Renilla* luciferase gene, luminescence in transfected CHO cells was detected visually without the aid of instrumentation.

In addition, the synthetic genes may be used to express fusion proteins, for example fusions with secretion leader sequences or cellular localization sequences, to study transcription in difficult-to-transfect cells such as primary cells, and/or to improve the analysis of regulatory pathways and genetic elements. Other uses include, but are not limited to, the detection of rare events that require extreme sensitivity (e.g., studying RNA recoding), use with IRES, to improve the efficiency of *in vitro* translation or *in vitro* transcription-translation coupled systems such as TNT (Promega Corp., Madison, WI), study of reporters optimized to different host organisms (e.g., plants, fungus, and the like), use of multiple genes as co-reporters to monitor drug toxicity, as reporter molecules in multiwell assays, and as reporter molecules in drug screening with the advantage of minimizing possible interference of reporter signal by different signal transduction pathways and other regulatory mechanisms.

Additionally, uses for the nucleic acid molecules of the invention include fluorescence activated cell sorting (FACS), fluorescent microscopy, to detect and/or measure the level of gene expression *in vitro* and *in vivo*, (e.g., to determine promoter strength), subcellular localization or targeting (fusion protein), as a marker, in calibration, in a kit, (e.g., for dual assays), for *in vivo* imaging, to analyze regulatory pathways and genetic elements, and in multi-well formats.

With respect to synthetic DNA encoding luciferases, the use of synthetic click beetle luciferases provides advantages such as the measurement of dual reporters. As *Renilla* luciferase is better suited for *in vivo* imaging (because it does not depend on ATP or Mg^{2+} for reaction, unlike firefly luciferase, and because coelenterazine is more permeable to the cell membrane than luciferin), the synthetic *Renilla* luciferase gene can be employed *in vivo*. Further, the synthetic *Renilla* luciferase has improved fidelity and sensitivity in dual luciferase assays, e.g., for biological analysis or in drug screening platform.

Demonstration of the Invention Using Luciferase Genes

The reporter genes for click beetle luciferase and *Renilla* luciferase were used to demonstrate the invention because the reaction catalyzed by the protein they encode are significantly easier to quantify than the product of most genes. However, for the purposes of demonstrating the present invention they represent genes in general.

Although the click beetle luciferase and *Renilla* luciferase genes share the name "luciferase", this should not be interpreted to mean that they originate from the same family of genes. The two luciferase proteins are evolutionarily distinct; they have fundamentally different traits and physical structures, they use vastly different substrates (Figure 17), and they evolved from completely different families of genes. The click beetle luciferase is 61 kD in size, uses luciferin as a substrate and evolved from the CoA synthetases. The *Renilla* luciferase originates from the sea pansy *Renilla Reniformis*, is 35 kD in size, uses coelenterazine as a substrate and evolved from the αβ hydrolases. The only shared trait of these two enzymes is that the reaction they catalyze results in light output. They are no more similar for resulting in light output than any other two enzymes would be, for example, simply because the reaction they catalyze results in heat.

Bioluminescence is the light produced in certain organisms as a result of luciferase-mediated oxidation reactions. The luciferase genes, e.g., the genes from luminous beetles, sea pansy, and, in particular, the luciferase from *Photinus pyralis* (the common firefly of North America), are currently the most popular luminescent reporter genes. Reference is made to Bronstein et al. (1994) for a review of luminescent reporter gene assays and to Wood (1995) for a review of the evolution of beetle bioluminescence. See Figure 17 for an illustration of the reactions catalyzed by each of firefly and click beetle luciferases (17A) and Renilla luciferase (17B).

Firefly luciferase and *Renilla* luciferase are highly valuable as genetic reporters due to the convenience, sensitivity and linear range of the luminescence assay. Today, luciferase is used in virtually every type of experimental biological system, including, but not limited to, prokaryotic and eukaryotic cell culture, transgenic plants and animals, and cell-free expression systems. The firefly luciferase enzyme is derived from a specific North American beetle, *Photinus pyralis*. The firefly luciferase enzyme and the click beetle luciferase enzyme are monomeric proteins (61 kDa) which generate light through monooxygenation of beetle luciferin utilizing ATP and O_2 (Figure 17A). The *Renilla* luciferase is derived from the sea pansy *Renilla reniformis*. The *Renilla* luciferase enzyme is a 36 kDa monomeric protein that utilizes O_2 and coelenterazine to generate light (Figure 17B).

The gene encoding firefly luciferase was cloned from *Photinus pyralis*, and demonstrated to produce active enzyme in *E. coli* (de Wet et al., 1987). The cDNA encoding firefly luciferase (*luc*) continues to gain favor as the gene of choice for reporting genetic activity in animal, plant and microbial cells. The firefly luciferase reaction, modified by the addition of CoA to produce persistent light emission, provides an extremely sensitive and rapid *in vitro* assay for quantifying firefly luciferase expression in small samples of transfected cells or tissues.

To use firefly luciferase or click beetle luciferase as a genetic reporter, extracts of cells expressing the luciferase are mixed with substrates (beetle luciferin, Mg^{2+} ATP, and O_2), and luminescence is measured immediately. The assay is very rapid and sensitive, providing gene expression data with little effort. The conventional firefly luciferase assay has been further improved by including coenzyme A in the assay reagent to yield greater enzyme turnover and thus greater luminescence intensity (Promega Luciferase Assay Reagent, Cat.# E1500, Promega Corporation, Madison, Wis.). Using this reagent, luciferase activity can be readily measured in luminometers or scintillation counters. Firefly and click beetle luciferase activity can also be detected in living cells in culture by adding luciferin to the growth medium. This *in situ* luminescence relies on the ability of beetle luciferin to diffuse through cellular and peroxisomal membranes and on the intracellular availability of ATP and O_2 in the cytosol and peroxisome.

Further, although reporter genes are widely used to measure transcription events, their utility can be limited by the fidelity and efficiency of reporter expression. For example, in U.S. Patent No. 5,670,356, a firefly luciferase gene (referred to as luc+) was modified to improve the level of luciferase expression. While a higher level of expression was observed, it was not determined that higher expression had improved regulatory control.

The invention will be further described by the following nonlimiting examples.

Example 1

Synthetic Click Beetle (RD and GR) Luciferase Nucleic Acid Molecules

Luc*Ppl*YG is a wild-type click beetle luciferase that emits yellow-green luminescence (Wood, 1989). A mutant of Luc*Ppl*YG named YG#81-6G01 was envisioned. YG#81-6G01 lacks a peroxisome targeting signal, has a lower K_M for luciferin and ATP, has increased signal stability and increased temperature stability when compared to the wild type (PCT/WO9914336). YG #81-6G01 was mutated to emit green luminescence by changing Ala at position 224 to Val (A224V is a green-shifting mutation), or to emit red luminescence by simultaneously introducing the amino acid substitutions A224H, S247H, N346I, and H348Q (red-shifting mutation set) (PCT/WO9518853)

Using YG #81-6G01 as a parent gene, two synthetic gene sequences were designed. One codes for a luciferase emitting green luminescence (GR) and one for a luciferase emitting red luminescence (RD). Both genes were designed to 1) have optimized codon usage for expression in mammalian cells, 2) have a reduced number of transcriptional regulatory sites including mammalian transcription factor binding sites, splice sites, poly(A) addition sites and promoters, as well as prokaryotic (*E. coli*) regulatory sites, 3) be devoid of unwanted restriction sites, e.g., those which are likely to interfere with standard cloning procedures, and 4) have a low DNA sequence identity compared to each other in order to minimize genetic rearrangements when both are present inside the same cell. In addition, desired sequences, e.g., a Kozak sequence or restriction enzyme recognition sites, may be identified and introduced.

Not all design criteria could be met equally well at the same time. The following priority was established for reduction of transcriptional regulatory sites: elimination of transcription factor (TF) binding sites received the highest priority, followed by elimination of splice sites and poly(A) addition sites, and finally prokaryotic regulatory sites. When removing regulatory sites, the strategy was to work from the lesser important to the most important to ensure that the most important changes were made last. Then the sequence was rechecked for the appearance of new lower priority sites and additional changes made as needed. Thus, the process for designing the synthetic GR and RD gene sequences, using computer programs described herein, involved 5 optionally iterative steps that are detailed below

1. Optimized codon usage and changed A224V to create <u>GRver1</u>, separately changed A224H, S247H, H348Q and N346I to create <u>RDver1.</u> These particular amino acid changes were maintained throughout all subsequent manipulations to the sequence.

2. Removed undesired restriction sites, prokaryotic regulatory sites, splice sites, poly(A) sites thereby creating GRver2 and RDver2.
3. Removed transcription factor binding sites (first pass) and removed any newly created undesired sites as listed in step 2 above thereby creating GRver3 and RDver3.
4. Removed transcription factor binding sites created by step 3 above (second pass) and removed any newly created undesired sites as listed in step 2 above thereby creating GRver4 and RDver4.
5. Removed transcription factor binding sites created by step 4 above (third Pass) and confirmed absence of sites listed in step 2 above thereby creating GRver5 and RDver5.
6. Constructed the actual genes by PCR using synthetic oligonucleotides corresponding to fragments of GRver5 and RDver5 designed sequences (Figures 6 and 10) thereby creating GR6 and RD7. GR6, upon sequencing was found to have the serine residue at amino acid position 49 mutated to an asparagine and the proline at amino acid position 230 mutated to a serine (S49N, P230S). RD7, upon sequencing was found to have the histidine at amino acid position 36 mutated to a tyrosine (H36Y). These changes occurred during the PCR process.
7. The mutations described in step 6 above (S49N, P230S for GR6 and H36Y for RD7) were reversed to create GRver5.1 and RDver5.1.
8. RDver5.1 was further modified by changing the arginine codon at position 351 to a glycine codon (R351G) thereby creating RDver5.2 with improved spectral properties compared to RDver5.1.
9. RDver5.2 was further mutated to increase luminescence intensity thereby creating RD156-1H9 which encodes four additional amino acid changes (M2I, S349T, K488T, E538V) and three silent single base changes (SEQ ID NO:18).

1. Optimize codon usage and introduce mutations determining luminescence color

The starting gene sequence for this design step was YG #81-6G01 (SEQ ID NO:2).

a) Optimize codon usage:

The strategy was to adapt the codon usage for optimal expression in human cells and at the same time to avoid *E. coli* low-usage codons. Based on these requirements, the best two codons for expression in human cells for all amino acids with more than two codons were selected (see Wada et al., 1990). In the selection of codon pairs for amino acids with six codons, the selection was biased towards pairs that have the largest number of mismatched bases to allow design of GR and RD genes with minimum sequence identity (codon distinction):

Arg: CGC/CGT Leu: CTG/TTG Ser: TCT/AGC
Thr: ACC/ACT Pro: CCA/CCT Ala: GCC/GCT
Gly: GGC/GGT Val: GTC/GTG Ile: ATC/ATT

Based on this selection of codons, two gene sequences encoding the YG#81-6G01 luciferase protein sequence were computer generated. The two genes were designed to have minimum DNA sequence identity and at the same time closely similar codon usage. To achieve this, each codon in the two genes was replaced by a codon from the limited list described above in an alternating fashion (e.g., Arg$_{(n)}$ is CGC in gene 1 and CGT in gene 2, Arg$_{(n+1)}$ is CGT in gene 1 and CGC in gene 2).

For subsequent steps in the design process it was anticipated that changes had to be made to this limited optimal codon selection in order to meet other design criteria, however, the following low-usage codons in mammalian cells were not used unless needed to meet criteria of higher priority:

Arg: CGA Leu: CTA Ser: TCG
Pro: CCG Val: GTA Ile: ATA

Also, the following low-usage codons in *E. coli* were avoided when reasonable (note that 3 of these match the low-usage list for mammalian cells):

Arg: CGA/CGG/AGA/AGG
Leu: CTA Pro: CCC Ile: ATA

b) Introduce mutations determining luminescence color:

Into one of the two codon-optimized gene sequences was introduced the single green-shifting mutation and into the other were introduced the 4 red-shifting mutations as described above.

The two output sequences from this first design step were named GRver1 (version 1 GR) and RDver1 (version 1 RD). Their DNA sequences are 63% identical (594 mismatches), while the proteins they encode differ only by the 4 amino acids that determine luminescence color (see Figures 2 and 3 for an alignment of the DNA and protein sequences).

Tables 1 and 2 show, as an example, the codon usage for valine and leucine in human genes, the parent gene YG#81-6G01, the codon-optimized synthetic genes GRver1 and RDver1, as well as the final versions of the synthetic genes after completion of step 5 in the design process (GRver5 and RDver5). For a complete summary of the codon changes, see Figures 4 and 5.

Table 1: Valine

Codon	Human	Parent	GR ver1	RD ver1	GR ver5	RD ver5
GTA	4	13	0	0	1	1
GTC	**13**	4	**25**	**24**	**21**	**26**
GTG	**24**	**12**	**25**	**25**	**25**	**17**
GTT	9	20	0	0	3	5

Table 2: Leucine

Codon	Human	Parent	GR ver1	RD ver1	GR ver5	RD ver5
CTA	3	5	0	0	0	.0
CTC	12	4	0	1	12	11
CTG	**24**	**4**	**28**	**27**	**19**	**18**
CTT	6	12	0	0	1	1
TTA	3	17	0	0	0	0
TTG	**6**	**13**	**27**	**27**	**23**	**25**

2. Remove undesired restriction sites, prokaryotic regulatory sites, splice sites and poly(A) addition sites

The starting gene sequences for this design step were GRver1 and RDver1.

a) Remove undesired restriction sites:

To check for the presence and location of undesired restriction sites, the sequences of both synthetic genes were compared against a database of restriction enzyme recognition sequences (REBASE ver.712, http://www.neb.com/rebase) using standard sequence analysis software (GenePro ver 6.10, Riverside Scientific Ent.).

Specifically, the following restriction enzymes were classified as undesired:

- *Bam*H I, *Xho* I, *Sfi* I, *Kpn* I, *Sac* I, *Mlu* I, *Nhe* I, *Sma* I, *Xho* I, *Bgl* II, *Hind* III, *Nco* I, *Nar* I, *Xba* I, *Hpa* I, *Sal* I,
- other cloning sites commonly used: *Eco*R I , *Eco*R V, *Cla* I,
- eight-base cutters (commonly used for complex constructs),
- *Bst*E II (to allow N-terminal fusions),
- *Xcm* I (can generate A/T overhang used for T-vector cloning).

To eliminate undesired restriction sites when found in a synthetic gene, one or more codons of the synthetic gene sequence were altered in accordance with the codon optimization guidelines described in 1a above.

b) Remove prokaryotic (*E. coli*) regulatory sequences:

To check for the presence and location of prokaryotic regulatory sequences, the sequences of both synthetic genes were searched for the presence of the following consensus sequences using standard sequence analysis software (GenePro):

- TATAAT (-10 Pribnow box of promoter)
- AGGA or GGAG (ribosome binding site; only considered if paired with a methionine codon 12 or fewer bases downstream).

To eliminate such regulatory sequences when found in a synthetic gene, one or more codons of the synthetic gene at sequence were altered in accordance with the codon optimization guidelines described in 1a above.

c) Remove splice sites:

To check for the presence and location of splice sites, the DNA strand corresponding to the primary RNA transcript of each synthetic gene was searched for the presence of the following consensus sequences (see Watson et al., 1983) using standard sequence analysis software (GenePro):

- splice donor site: AG | GTRAGT (exon | intron), the search was performed for AGGTRAG and the lower stringency GGTRAGT;
- splice acceptor site: $(Y)_n$NCAG | G (intron | exon), the search was performed with n = 1.

To eliminate splice sites found in a synthetic gene, one or more codons of the synthetic gene sequence were altered in accordance with the codon optimization guidelines described in 1a above. Splice acceptor sites were generally difficult to eliminate in one gene without introducing them into the other gene because they tended to contain one of the two only Gln codons (CAG); they were removed by placing the Gln codon CAA in both genes at the expense of a slightly increased sequence identity between the two genes.

d) Remove poly(A) addition sites:

To check for the presence and location of poly(A) addition sites, the sequences of both synthetic genes were searched for the presence of the following consensus sequence using standard sequence analysis software (GenePro):

- AATAAA.

To eliminate each poly(A) addition site found in a synthetic gene, one or more codons of the synthetic gene sequence were altered in accordance with the codon optimization guidelines described in 1a above. The two output sequences from this second design step were named GRver2 and RDver2. Their DNA sequences are 63% identical (590 mismatches) (Figs. 2 and 3).

3. Remove transcription factor (TF) binding sites, then repeat steps 2 a-d

The starting gene sequences for this design step were GRver2 and RDver2. To check for the presence, location and identity of potential TF binding sites, the sequences of both synthetic genes were used as query sequences to search a database of transcription factor binding sites (TRANSFAC v3.2). The TRANSFAC database (http://transfac.gbf.de/TRANSFAC/index:html) holds information on gene regulatory DNA sequences (TF binding sites) and proteins (TFs) that bind to and act through them. The SITE table of TRANSFAC Release 3.2 contains 4,401 entries of individual (putative) TF binding sites (including TF binding sites in eukaryotic genes, in artificial sequences resulting from mutagenesis studies and *in vitro* selection procedures based on random oligonucleotide mixtures or specific theoretical considerations, and consensus binding sequences (from Faisst and Meyer, 1992)).

The software tool used to locate and display these TF binding sites in the synthetic gene sequences was TESS (Transcription Element Search Software, http://agave.humgen.upenn.edu/tess/index.html). The filtered string-based search option was used with the following user-defined search parameters:

- Factor Selection Attribute: Organism Classification
- Search Pattern: Mammalia
- Max. Allowable Mismatch %: 0
- Min. element length: 5
- Min. log-likelihood: 10

This parameter selection specifies that only mammalian TF binding sites (approximately 1,400 of the 4,401 entries in the database) that are at least 5 bases long will be included in the search. It further specifies that only TF binding sites that have a perfect match in the query sequence and a minimum log likelihood (LLH) score of 10 will be reported. The LLH scoring method assigns 2 to an unambiguous match, 1 to a partially ambiguous match (e.g., A or T match W) and 0 to a match against 'N'. For example, a search with parameters specified above would result in a "hit" (positive result or match) for TATAA (SEQ ID NO:240) (LLH = 10), STRATG (SEQ ID NO:241) (LLH = 10), and MTTNCNNMA (SEQ ID NO:242) (LLH = 10) but not for TRATG (SEQ ID NO: 243) (LLH = 9) if these four TF binding sites were present in the query sequence. A lower stringency test was performed at the end of the design process to re-evaluate the search parameters.

When TESS was tested with a mock query sequence containing known TF binding sites it was found that the program was unable to report matches to sites ending with the 3′ end of the query sequence. Thus, an extra nucleotide was added to the 3' end of all query sequences to eliminate this problem.

The first search for TF binding sites using the parameters described above found about 100 transcription factor binding sites (hits) for each of the two synthetic genes (GRver2 and RDver2). All sites were eliminated by changing one or more codons of the synthetic gene sequences in accordance with the codon optimization guidelines described in 1a above. However, it was expected that some these changes created new TF binding sites, other regulatory sites, and new restriction sites. Thus, steps 2 a-d were repeated as described, and 4 new restriction sites and 2 new splice sites were removed. The two output sequences from this third design step were named GRver3 and RDver3. Their DNA sequences are 66% identical (541 mismatches) (Figs. 2 and 3).

4. Remove new transcription factor (TF) binding sites, then repeat steps 2 a-d

The starting gene sequences for this design step were GRver3 and RDver3. This fourth step is an iteration of the process described in step 3. The search for newly introduced TF binding sites yielded about 50 hits for each of the two synthetic genes. All sites were eliminated by changing one or more codons of the synthetic gene sequences in general accordance with the codon optimization guidelines described in 1a above. However, more high to medium usage codons were used to allow elimination of all TF binding sites. The lowest priority was placed on maintaining low sequence identity between the GR and RD genes. Then steps 2 a-d were repeated as described. The two output sequences from this fourth design step were named GRver4 and RDver4. Their DNA sequences are 68% identical (506 mismatches) (Figs 2 and 3).

5. Remove new transcription factor (TF) binding sites, then repeat steps 2 a-d

The starting gene sequences for this design step were GRver4 and RDver4. This fifth step is another iteration of the process described in step 3 above. The search for new TF binding sites introduced in step 4 yielded about 20 hits for each of the two synthetic genes. All sites were eliminated by changing one or more codons of the synthetic gene sequences in general accordance with the codon optimization guidelines described in 1a above. However, more high to medium usage codons were used (these are all considered "preferred") to allow elimination of all TF binding sites. The lowest priority was placed on maintaining low sequence identity between the GR and RD genes. Then steps 2 a-d were repeated as described. Only one acceptor splice site could not be eliminated. As a final step the absence of all TF binding sites in both genes as specified in step 3 was confirmed. The two output sequences from this fifth and last design step were named GRver5 and RDver5. Their DNA sequences are 69% identical (504 mismatches) (Figs. 2 and 3).

Additional evaluation of GRver5 and RDver5

a) Use lower stringency parameters for TESS:

The search for TF binding sites was repeated as described in step 3 above, but with even less stringent user-defined parameters:

- setting LLH to 9 instead of 10 did not result in new hits;
- setting LLH to 0 through 8 (incl.) resulted in hits for two additional sites, MAMAG (22 hits) and CTKTK (24 hits);
- setting LLH to 8 and the minimum element length to 4, the search yielded (in addition to the two sites above) different 4-base sites for AP-1, NF-1, and c-Myb that are shortened versions of their longer respective consensus sites which were eliminated in steps 3-5 above.

It was not realistic to attempt complete elimination of these sites without introduction of new sites, so no further changes were made.

b) Search different database:

The Eukaryotic Promoter Database (release 45) contains information about reliably mapped transcription start sites (1253 sequences) of eukaryotic genes. This database was searched using BLASTN 1.4.11 with default parameters (optimized to find nearly identical sequences rapidly; see Altschul et al, 1990) at the National

Center for Biotechnology Information site (http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST). To test this approach, a portion of pGL3-Control vector sequence containing the SV40 promoter and enhancer was used as a query sequence, yielding the expected hits to SV40 sequences. No hits were found when using the two synthetic genes as query sequences.

Summary of GRver5 and RDver5 synthetic gene properties

Both genes, which at this stage were still only "virtual" sequences in the computer, have a codon usage that strongly favors mammalian high-usage codons and minimizes mammalian and *E. coli* low-usage codons. Figure 4 shows a summary of the codon usage of the parent gene and the various synthetic gene versions.

Both genes are also completely devoid of eukaryotic TF binding sites consisting of more than four unambiguous bases, donor and acceptor splice sites (one exception: GRver5 contains one splice acceptor site), poly(A) addition sites, specific prokaryotic (*E. coli*) regulatory sequences, and undesired restriction sites.

The gene sequence identity between GRver5 and RDver5 is only 69% (504 base mismatches) while their encoded proteins are 99% identical (4 amino acid mismatches), see Figures 2 and 3. Their identity with the parent sequence YG#81-6G1 is 74% (GRver5) and 73% (RDver5), see Figure 2. Their base composition is 49.9% GC (GRver5) and 49.5% GC (RDver5), compared to 40.2% GC for the parent YG#81-6G01.

Construction of synthetic genes

The two synthetic genes were constructed by assembly from synthetic oligonucleotides in a thermocycler followed by PCR amplification of the full-length genes (similar to Stemmer et al. (1995) Gene. 164, pp. 49-53). Unintended mutations that interfered with the design goals of the synthetic genes were corrected.

a) Design of synthetic oligonucleotides:

The synthetic oligonucleotides were mostly 40mers that collectively code for both complete strands of each designed gene (1,626 bp) plus flanking regions needed for cloning (1,950 bp total for each gene; Figure 6). The 5' and 3' boundaries of all oligonucleotides specifying one strand were generally placed in a manner to give an average offset/overlap of 20 bases relative to the boundaries of the oligonucleotides specifying the opposite strand.

The ends of the flanking regions of both genes matched the ends of the amplification primers (pRAMtailup: 5′-gtactgagacgacgccagcccaagcttaggcctgagtg SEQ ID NO:229, and pRAMtaildn: 5′-ggcatgagcgtgaactgactgaactagcggccgccgag SEQ ID NO:230) to allow cloning of the genes into our *E. coli* expression vector pRAM (WO99/14336).

A total of 183 oligonucleotides were designed (Figure 6): fifteen oligonucleotides that collectively encode the upstream and downstream flanking sequences (identical for both genes; SEQ ID NOs: 35-49) and 168 oligonucleotides (4 x 42) that encode both strands of the two genes (SEQ ID NOs: 50-217).

All 183 oligonucleotides were run through the hairpin analysis of the OLIGO software (OLIGO 4.0 Primer Analysis Software © 1989-1991 by Wojciech Rychlik) to identify potentially detrimental intra-molecular loop formation. The guidelines for evaluating the analysis results were set according to recommendations of Dr. Sims (Sigma-Genosys Custom Gene Synthesis Department): oligos forming hairpins with $\Delta G < -10$ have to be avoided, those forming hairpins with $\Delta G \leq -7$ involving the 3' end of the oligonucleotide should also be avoided, while those with an overall $\Delta G \leq -5$ should not pose a problem for this application. The analysis identified 23 oligonucleotides able to form hairpins with a ΔG between -7.1 and -4.9. Of these, 5 had blocked or nearly blocked 3′ ends (0-3 free bases) and were re-designed by removing 1-4 bases at their 3' end and adding it to the adjacent oligonucleotide.

The 40mer oligonucleotide covering the sequence complementary to the poly(A) tail had a very low complexity 3′ end (13 consecutive T bases). An additional 40mer was designed with a high complexity 3′ end but a consequently reduced overlap with one of its complementary oligonucleotides (11 instead of 20 bases) on the opposite strand.

Even though the oligos were designed for use in a thermocycler-based assembly reaction, they could also be used in a ligation-based protocol for gene construction. In this approach, the oligonucleotides are annealed in a pairwise fashion and the resulting short double-stranded fragments are ligated using the sticky overhangs. However, this would require that all oligonucleotides be phosphorylated.

b) Gene assembly and amplification

In a first step, each of the two synthetic genes was assembled in a separate reaction from 98 oligonucleotides. The total volume for each reaction was 50 µl:

0.5 µM oligonucleotides (= 0.25 pmoles of each oligo)
1.0 U *Taq* DNA polymerase
0.02 U *Pfu* DNA polymerase
2 mM $MgCl_2$
0.2 mM dNTPs (each)
0.1% gelatin
Cycling conditions: (94°C for 30 seconds, 52°C for 30 seconds, and 72°C for 30 seconds) x 55 cycles.

In a second step, each assembled synthetic gene was amplified in a separate reaction. The total volume for each reaction was 50 µl:

2.5 1 assembly reaction
5.0 U *Taq* DNA polymerase
0.1 U *Pfu* DNA polymerase
1 M each primer (pRAMtailup, pRAMtaildn)
2 mM $MgCl_2$
0.2 mM dNTPs (each)
Cycling conditions: (94°C for 20 seconds, 65°C for 60 seconds, 72°C for 3 minutes) x 30 cycles.

The assembled and amplified genes were subcloned into the pRAM vector and expressed in *E. coli*, yielding 1-2% luminescent GR or RD clones. Five GR and five RD clones were isolated and analyzed further. Of the five GR clones, three had the correct insert size, of which one was weakly luminescent and one had an altered restriction pattern. Of the five RD clones, two had the correct size insert with an altered restriction pattern and one of those was weakly luminescent. Overall, the analysis indicated the presence of a large number of mutations in the genes, most likely the result of errors introduced in the assembly and amplification reactions.

c) Corrective assembly and amplification

To remove the large number of mutations present in the full-length synthetic genes we performed an additional assembly and amplification reaction for each gene using the proof-reading DNA polymerase *Tli*. The assembly reaction contained, in addition to the 98 GR or RD oligonucleotides, a small amount of DNA from the corresponding full-length clones with mutations described above. This allows the oligos to correct mutations present in the templates.

The following assembly reaction was performed for each of the synthetic genes. The total volume for each reaction was 50 μl:

0.5 μM oligonucleotides (= 0.25 pmoles of each oligo)
0.016 pmol plasmid (mix of clones with correct insert size)
2.5 U *Tli* DNA polymerase
2 mM $MgCl_2$
0.2 mM dNTPs (each)
0.1% gelatin
Cycling conditions: 94°C for 30 seconds, then (94°C for 30 seconds, 52°C for 30 seconds, 72°C for 30 seconds) for 55 cycles, then 72°C for 5 minutes.

The following amplification reaction was performed on each of the assembly reactions. The total volume for each amplification reaction was 50 μl:

1-5 μl of assembly reaction

40 pmol each primer (pRAMtailup, pRAMtaildn)

2.5 U *Tli* DNA polymerase

2 mM $MgCl_2$

0.2 mM dNTPs (each)

Cycling conditions: 94°C for 30 seconds, then (94°C for 20 seconds, 65°C for 60 seconds and 72°C for 3 minutes) for 30 cycles, then 72°C for 5 minutes.

The genes obtained from the corrective assembly and amplification step were subcloned into the pRAM vector and expressed in *E. coli*, yielding 75% luminescent GR or RD clones. Forty-four GR and 44 RD clones were analyzed with our screening robot (WO99/14336). The six best GR and RD clones were manually analyzed and one best GR and RD clone was selected (GR6 and RD7). Sequence analysis of GR6 revealed two point mutations in the coding region, both of which resulted in an amino acid substitution (S49N and P230S). Sequence analysis of RD7 revealed three point mutations in the coding region, one of which resulted in an amino acid substitution (H36Y). It was confirmed that none of the silent point mutations introduced any regulatory or restriction sites conflicting with the overall design criteria for the synthetic genes.

d) Reversal of unintended amino acid substitutions

The unintended amino acid substitutions present in the GR6 and RD7 synthetic genes were reversed by site-directed mutagenesis to match the GRver5 and RDver5 designed sequences, thereby creating GRver5.1 and RDver5.1. The DNA sequences of the mutated regions were confirmed by sequence analysis.

e) Improve spectral properties

The RDver5.1 gene was further modified to improve its spectral properties by introducing an amino change (R351G), thereby creating RDver5.2

<u>pGL3 vectors with RD and GR genes</u>

The parent click beetle luciferase YG#81-6G1 ("YG"), and the synthetic click beetle luciferase genes GRver5.1 ("GR"), RDver5.2 ("RD"), and RD156-1H9 were cloned into the four pGL3 reporter vectors (Promega Corp.):

- pGL3-Basic = no promoter, no enhancer
- pGL3-Control = SV40 promoter, SV40 enhancer
- pGL3-Enhancer = SV40 enhancer (3' to luciferase coding sequences)
- pGL3-Promoter = SV40 promoter.

The primers employed in the assembly of GR and RD synthetic genes facilitated the cloning of those genes into pRAM vectors. To introduce the genes into pGL3 vectors (Promega Corp., Madison, WI) for analysis in mammalian cells, each gene in a pRAM vector (pRAM RDver5.1, pRAM GRver5.1, and pRAM RD156-1H9) was amplified to introduce an *Nco* I site at the 5′ end and an *Xba* I site at the 3′ end of the gene. The primers for pRAM RDver5.1 and pRAM GRver5.1 were:

GR→5′ GGA TCC CAT GGT GAA GCG TGA GAA 3′ (SEQ ID NO:231) or

RD→5′ GGA TCC CAT GGT GAA ACG CGA 3′ (SEQ ID NO:232) and

5′ CTA GCT TTT TTT TCT AGA TAA TCA TGA AGA C 3’ (SEQ ID NO:233)

The primers for pRAM RD156-1H9 were:

5′ GCG TAG CCA TGG TAA AGC GTG AGA AAA ATG TC 3′ (SEQ ID NO: 295) and

5′ CCG ACT CTA GAT TAC TAA CCG CCG GCC TTC ACC 3′ (SEQ ID NO: 296)

The PCR included:

100 ng DNA plasmid

1 μM primer upstream

1 μM primer downstream

0.2 mM dNTPs

1X buffer (Promega Corp.)

5 units *Pfu* DNA polymerase (Promega Corp.)

Sterile nanopure H_2O to 50 μl

The cycling parameters were: 94°C for 5 minutes; (94°C for 30 seconds; 55°C for 1 minute; and 72°C for 3 minutes) x 15 cycles. The purified PCR product was digested with *Nco* I and *Xba* I, ligated with pGL3-control that was also digested with *Nco* I and *Xba* I, and the ligated products introduced to *E. coli*. To insert the luciferase genes into the other pGL3 reporter vectors (basic, promoter and enhancer), the pGL3-control vectors containing each of the luciferase genes was digested with *Nco* I and *Xba* I, ligated with other pGL3 vectors that also were digested with *Nco* I and *Xba* I, and the ligated products introduced to *E. coli*. Note that the polypeptide encoded by GRver5.1 and RDver5.1 (and RD156-1H9, see below) nucleic acid sequences in pGL3 vectors has an amino acid substitution at position 2 to valine as a result of the *Nco* I site at the initiation codon in the oligonucleotide.

Because of internal *Nco* I and *Xba* I sites, the native gene in YG #81-6G01 was amplified from a *Hind* III site upstream to a *Hpa* I site downstream of the coding region and which included flanking sequences found in the GR and RD clones. The upstream primer (5′-CAA AAA GCT TGG CAT TCC GGT ACT GTT GGT AAA GCC ACC ATG GTG AAG CGA GAG- 3′; SEQ ID NO:234) and a downstream primer (5′- CAA TTG TTG TTG TTA ACT TGT TTA TT -3′; SEQ ID NO:235) were mixed with YG#81-6G01 and amplified using the PCR conditions above. The purified PCR product was digested with *Nco* I and *Xba* I, ligated with pGL3-control that was also digested with *Hind* III and *Hpa* I, and the ligated products introduced into *E. coli*. To insert YG#81-6G01 into the other pGL3 reporter vectors (basic, promoter and enhancer), the pGL3-control vectors containing YG#81-6G01 were digested with *Nco* I and *Xba* I, ligated with the other pGL3 vectors that also were digested with *Nco* I and *Xba* I, and the ligated products introduced to *E. coli*. Note that the clone of YG#81-6G01 in the pGL3 vectors has a C instead of an A at base 786, which yields a change in the amino acid sequence at residue 262 from Phe to Leu (Figure 2 shows the sequence of YG#81-6G01 prior to introduction into pGL3 vectors). To determine whether the altered amino acid at position 262 affected the enzyme biochemistry, the clone of YG#81-6G01 was mutated to resemble the original sequence. Both clones were then tested for expression in *E. coli*, physical stability, substrate binding, and luminescence output kinetics. No significant differences were found.

Partially purified enzymes expressed from the synthetic genes and the parent gene were employed to determine Km for luciferin and ATP (see Table 3).

Table 3

Enzyme	K_M (LH_2)	K_M (ATP)
YG parent	2 μM	17 μM
GR	1.3 μM	25 μM
RD	24.5 μM	46 μM

In vitro eukaryotic transcription/translation reactions were also conducted using Promega's TNT T7 Quick system according to manufacturer's instructions. Luminescence levels were 1 to 37-fold and 1 to 77-fold higher (depending on the reaction time) for the synthetic GR and RD genes, respectively, compared to the parent gene (corrected for luminometer spectral sensitivity).

To test whether the synthetic click beetle luciferase genes and the wild type click beetle gene have improved expression in mammalian cells, each of the synthetic genes and the parent gene was cloned into a series of pGL3 vectors and introduced into CHO cells (Table 8). In all cases, the synthetic click beetle genes exhibited a higher expression than the native gene. Specifically, expression of the synthetic GR and RD genes was 1900-fold and 40-fold higher, respectively, than that of the parent (transfection efficiency normalized by comparison to native *Renilla* luciferase gene). Moreover, the data (basic versus control vector) show that the synthetic genes have reduced basal level transcription.

Further, in experiments with the enhancer vector where the percentage of activity in reference to the control is compared between the native and synthetic gene, the data showed that the synthetic genes have reduced risk of anomalous transcription characteristics. In particular, the parent gene appeared to contain one or more internal transcriptional regulatory sequences that are activated by the enhancer in the vector, and thus is not suitable as a reporter gene while the synthetic GR and RD genes showed a clean reporter response (transfection efficiency normalized by comparison to native *Renilla* luciferase gene). See Table 9.

The clone names and their corresponding SEQ ID numbers for nucleotide sequence and amino acid sequence are listed below in Table 4.

Table 4

Clone name	Luciferase Type	SEQ ID NO.	SEQ ID NO.
LUCPPLYG	Wild type YG Click Beetle	1	23
YG#81-6G01	Mutant YG Click Beetle	2	24
GRver1	Synthetic Green Click Beetle	3	25
GRver2	Synthetic Green Click Beetle	4	26
GRver3	Synthetic Green Click Beetle	5	27
GRver4	Synthetic Green Click Beetle	6	28
GRver5	Synthetic Green Click Beetle	7	29
GR6	Synthetic Green Click Beetle	8	30
GRver5.1	Synthetic Green Click Beetle	9	31
RDver1	Synthetic Red Click Beetle	10	32
RDver2	Synthetic Red Click Beetle	11	33
RDver3	Synthetic Red Click Beetle	12	34
RDver4	Synthetic Red Click Beetle	13	218
RDver5	Synthetic Red Click Beetle	14	219
RD7	Synthetic Red Click Beetle	15	220
RDver5.1	Synthetic Red Click Beetle	16	221
RDver5.2	Synthetic Red Click Beetle	17	222
RD156-1H9	Synthetic Red Click Beetle	18	223
RELLUC	Wild type *Renilla*	19	224
Rlucver1	Synthetic *Renilla*	20	225
Rlucver2	Synthetic *Renilla*	21	226

Rluc-final	Synthetic *Renilla*	22	227

Example 2

Evolution of the RD luciferase gene

RDver5.2 was mutated to increase its luminescence intensity, thereby creating RD156-1H9 which carries four additional amino acid changes (M2I, S349T, K488T, E538V) and three silent point mutations (SEQ ID NO:18).

a) Site-directed mutagenesis:

The initial strategy was to use site-directed mutagenesis. There are four amino acid differences between the GR and RD synthetic genes with H348Q providing the greatest contribution to red color. Thus, this substitution may also cause structural changes in the protein that could lead to low light output. Optimization of positions near this area could increase light output. The following positions were selected for mutagenesis:

1. S344 (at the edge of the binding pocket for luciferin) – randomize this codon.
2. A245 (strictly conserved but closest to 348 and at the edge of the active site pocket) – randomize this codon.
3. I347 (not conserved, next to 348 in sequence) – mutate to hydrophobic amino acids only.
4. S349 (not conserved, next to 348 in sequence) – mutate to S, T, A, P only.

Oligonucleotides designed to mutate the above positions were used in a site-directed mutagenesis experiment (WO99/14336) and the resulting mutants were screened for luminescence intensity. There was little variation in light intensity and only about 25% were luminescent. For more detailed analysis, clones were picked and analyzed with the screening robot (PCT/WO9914336). None of the clones had a luminescence intensity (LI) higher than RDver5.2, but four of the clones had slightly lower composite Km for luciferin and ATP (Km).

b) Directed evolution:

Protocols and procedures used for the directed evolution are detailed in see PCT/WO9914336. DNA from the four clones with lower Km was combined and three libraries of random mutants were produced. The libraries were screened with the robot and clones with the highest LI values were selected. These clones were shuffled together and another robotic screen was completed with an incubation temperature of 46°C. The three clones with the highest LI values were RD156-0B4, RD156-1A5, and RD156-1H9.

c) Analysis:

The three clones with the highest LI values were selected for manual analysis to confirm that their luminescence intensity was higher than that of RDver5.2 and to ensure that their spectral properties were not compromised. One of the clones was slightly green-shifted, all others maintained the spectral properties of RDver5.2 (Table 5).

Table 5

Clone	Peak (nm)	Width (nm)
RD156-0B4	616	68
RD156-1A5	614	70
RD156-1H9	618	69
RDver5.2 (prep #1)	617	70
RDver5.2 (prep #2)	618	69

The Km values for luciferin and the luminescence intensity relative to RDver5.2 were determined for all three clones in several independent experiments. All cells samples were processed with CCLR lysis buffer (E1483, Promega Corp., Madison, WI) and diluted 1: 10 into buffer (25 mM HEPES pH 7.8, 5% glycerol, 1 mg/ml BSA, 150 mM NaCl). Table 7 summarizes the results (Lum: luminescence values were normalized to optical density; measurements for independent experiments are separated by forward slashes) from expression in bacterial cells. RD156-1H9, the clone with the highest luminescence intensity (5 to 10-fold increase) also has an about 2-fold higher Km for luciferin.

Table 6

Clone	Km Luciferin [μM]	Lum (normalized to RDver5.2)
RD156-0B4	8 /10	2.2 / 2.5
RD156-1A5	13 / 13	3.1 / 5.6
RD156-1H9	20 / 23 / 23	4 / 10.9 / 7.5
RDver5.2 (prep #1)	12 / 14 / 14	
RDver5.2 (prep #2)	40 / 50	
GRver5.1 (prep #1)	0.5	64
GRver5.1 (prep #2)	3	

Table 7 shows a comparison between the luminescence intensities of RD156-1H9, GRver5.1 and RDver5.2 normalized to GRver5.1 with and without correction for the spectral sensitivity of the luminometer photomultiplier tube. With correction, the luminescence intensity of clone RD156-1H9 was only about 2-fold lower than that of GRver5.1. The luciferin Km for clone RD156-1H9 is approximately 40-fold higher than GRver5.1. RD156-1H9 is thermostable at 50°C for at least 2 hours.

Table 7

Name	No Correction	With Correction
RDver5.2	0.016	0.06
GRver5.1	1.000	1.00
RD156-1H9	0.116	0.45

Tables 8 and 9 show a comparison of luciferase expression levels in CHO cells. Table 8 shows the expression levels only from the control vectors in comparison to the firefly luciferase gene (RLU = relative light units). Table 9 shows a comparison of the expression levels in all four pGL3 vectors calculated as a percent of the expression level in pGL3-control.

Table 8

Synthetic Click Beetle Gene Expression

Control vector	rlu
YG#81-6G01	177
GRver5.1	343,417
RDver5.1	7,161
RD156-1H9	20,802
FireFly	488,016

Table 9

Synthetic Click Beetle Gene Expression

Vector	Percent of control vector
YG-control	100
RD-control	100
GR-control	100
RD156-1H9 control	100
YG-basic	3.3
RD-basic	1.0
GR-basic	0.2
RD156-1H9 basic	0.3
YG-promoter	4.2
RD-promoter	15.1
GR-promoter	5.7
RD156-1H9 promoter	15.5
YG-enhancer	51.5
RD-enhancer	2.8
GR-enhancer	1.4
RD156-1H9 enhancer	0.3

Example 3

Synthetic *Renilla* Luciferase Nucleic Acid Molecule

The synthetic *Renilla* luciferase genes prepared include 1) an introduced Kozak sequence, 2) codon usage optimized for mammalian (human) expression, 3) a reduction or elimination of unwanted restriction sites, 4) removal of prokaryotic regulatory sites (ribosome binding site and TATA box), 5) removal of splice sites and poly(A) addition sites, and 6) a reduction or elimination of mammalian transcriptional factor binding sequences.

The process of computer-assisted design of synthetic *Renilla* luciferase genes by iterative rounds of codon optimization and removal of transcription factor binding sites and other regulatory sites as well as restriction sites can be described in three steps:

1. Using the wild type *Renilla* luciferase gene as the parent gene, codon usage was optimized, one amino acid was changed (T→A) to generate a Kozak consensus sequence, and undesired restriction sites were eliminated thereby creating synthetic gene Rlucver1.
2. Remove prokaryotic regulatory sites, splice sites, poly(A) sites and transcription factor (TF) binding sites (first pass). Then remove newly created TF binding sites. Then remove newly created undesired restriction enzyme sites, prokaryotic regulatory sites, splice sites, and poly(A) sites without introducing new TF binding sites. This thereby created Rlucver2.
3. Change 3 bases of Rlucver2 thereby creating Rluc-final.
4. The actual gene was then constructed from synthetic oligonucleotides corresponding to the Rluc-final designed sequence. All mutations resulting from the assembly or PCR process were corrected. This gene is Rluc-final (SEQ ID NO:22) and encodes the amino acid sequence of SEQ ID NO:227.

Codon Selection

Starting with the *Renilla reniformis* luciferase sequence in Genbank (Accession No. M63501, SEQ ID NO:19), codons were selected based on codon usage for optimal expression in human cells and to avoid *E. coli* low-usage codons. The best codon for expression in human cells (or the best two codons if found at a similar frequency) was chosen for all amino acids with more than one codon (Wada et al., 1990):

Arg: CGC	Lys: AAG

Leu: CTG	Asn: AAC
Ser: TCT/AGC	Gln: CAG
Thr: ACC	His: CAC
Pro: CCA/CCT	Glu: GAG
Ala: GCC	Asp: GAC
Gly: GGC	Tyr: TAC
Val: GTG	Cys: TGC
Ile: ATC/ATT	Phe: TTC

In cases where two codons were selected for one amino acid, they were used in an alternating fashion. To meet other criteria for the synthetic gene, the initial optimal codon selection was modified to some extent later. For example, introduction of a Kozak sequence required the use of GCT for Ala at amino acid position 2 (see below).

The following low-usage codons in mammalian cells were not used unless needed: Arg: CGA, CGU; Leu: CTA, UUA; Ser: TCG; Pro: CCG; Val: GTA; and Ile: ATA. The following low-usage codons in *E. coli* were also avoided when reasonable (note that 3 of these match the low-usage list for mammalian cells): Arg: CGA/CGG/AGA/AGG, Leu: CTA; Pro: CCC; Ile: ATA.

Introduction of Kozak Sequences

The Kozak sequence: 5′ aaccATGGCT 3′ (SEQ ID NO: 293) (the *Nco* I site is underlined, the coding region is shown in capital letters) was introduced to the synthetic *Renilla* luciferase gene. The introduction of the Kozak sequence changes the second amino acid from Thr to Ala (GCT).

Removal of undesired restriction sites

REBASE ver. 808 (updated August 1, 1998; Restriction Enzyme Database; www.neb.com/rebase) was employed to identify undesirable restriction sites as described in Example 1. The following undesired restriction sites (in addition to those described in Example 1) were removed according to the process described in Example 1: *Eco*ICR I, *Nde*I, *Nsi*I, *Sph*I, *Spe*I, *Xma*I, *Pst*I.

The version of *Renilla* luciferase (Rluc) which incorporates all these changes is Rlucver1.

Removal of prokaryotic (*E. coli*) regulatory sequences, splice sites, and poly(A) sites

The priority and process for eliminating transcription regulation sites was as described in Example 1.

Removal of TF binding sites

The same process, tools, and criteria were used as described in Example 1, however, the newer version 3.3 of the TRANSFAC database was employed.

After removing prokaryotic regulatory sequences, splice sites and poly(A) sites from Rlucver1, the first search for TF binding sites identified about 60 hits. All sites were eliminated with the exception of three that could not be removed without altering the amino acid sequence of the synthetic *Renilla* gene:

1. site at position 63 composed of two codons for W (TGGTGG), for CAC-binding protein T00076;
2. site at position 522 composed of codons for KMV (AAN ATG GTN), for myc-DF1 T00517;
3. site at position 885 composed of codons for EMG (GAR ATG GGN), for myc-DF1 T00517.

The subsequent second search for (newly introduced) TF binding sites yielded about 20 hits. All new sites were eliminated, leaving only the three sites described above. Finally, any newly introduced restriction sites, prokaryotic regulatory sequences, splice sites and poly(A) sites were removed without introducing new TF binding sites if possible.

Rlucver2 was obtained (SEQ ID Nos. 21 and 226).

As in Example 1, lower stringency search parameters were specified for the TESS filtered string search to further evaluate the synthetic *Renilla* gene.

With the LLH reduced from 10 to 9 and the minimum element length reduced from 5 to 4, the TESS filtered string search did not show any new hits.

When, in addition to the parameter changes listed above, the organism classification was expanded from "mammalia" to "chordata", the search yielded only four more TF binding sites. When the Min LLH was further reduced to between 8 and 0, the search showed two additional 5-base sites (MAMAG and CTKTK) which combined had four matches in Rlucver2, as well as several 4-base sites. Also as in Example 1, Rlucver2 was checked for hits to entries in the EPD (Eukaryotic Promoter Database, Release 45). Three hits were determined (one to Mus musculus promoter H-2L^d (Cell, 44, 261 (1986), one to Herpes Simplex Virus type 1 promoter b'g'2.7 kb, and one to Homo sapiens DHFR promoter (J. Mol. Biol., 176, 169 (1984)). However, no further changes were made to Rlucver2.

Summary of Properties for Rlucver2

- All 30 low usage codons were eliminated. The introduction of a Kozak sequence changed the second amino acid from Thr to Ala;
- base composition: 55.7% GC (*Renilla* wild-type parent gene: 36.5%);
- one undesired restriction site could not be eliminated: *Eco*R V at position 488;
- the synthetic gene had no prokaryotic promoter sequence but one potentially functional ribosome binding site (RBS) at positions 867-73 (about 13 bases upstream of a Met codon) could not be eliminated;
- all poly(A) addition sites were eliminated;
- splice sites: 2 donor splice sites could not be eliminated (both share the amino acid sequence MGK);
- TF sites: all sites with a consensus of >4 unambiguous bases were eliminated (about 280 TF binding sites were removed) with 3 exceptions due to the preference to avoid changes to the amino acid sequence.

Synthetic *Renilla* luciferase sequences are shown in Figures 7 and 8. A codon usage comparison is shown in Figure 9.

When introduced into pGL3, Rluc-final has a Kozak sequence (CACCATGGCT). The changes in Rluc-final relative to Rlucver2 were introduced during gene assembly. One change was at position 619, a C to an A, which eliminated a eukaryotic promoter sequence and reduced the stability of a hairpin structure in the corresponding oligonucleotide employed to assemble the gene. Other changes included a change from CGC to AGA at positions 218-220 (resulted in a better oligonucleotide for PCR).

<u>Gene Assembly Strategy</u>

The gene assembly protocol employed for the synthetic *Renilla* luciferase was similar to that described in Example 1. The oligonucleotides employed are shown in Figure 10.

Sense Strand primer:

5′ AACCATGGCTTCCAAGGTGTACGACCCCGAGCAACGCAAA 3′ (SEQ ID NO:236)

Anti-sense Strand primer:

5′ GCTCTAGAATTACTGCTCGTTCTTCAGCACGCGCTCCACG 3′ (SEQ ID NO:237)

The resulting synthetic gene fragment was cloned into a pRAM vector using *Nco* I and *Xba* I. Two clones having the correct size insert were sequenced. Four to six mutations were found in the synthetic gene from each clone. These mutations were fixed by site-directed mutagenesis (Gene Editor from Promega Corp., Madison, WI) and swapping the correct regions between these two genes. The corrected gene was confirmed by sequencing.

<u>Other Vectors</u>

To prepare an expression vector for the synthetic *Renilla* luciferase gene in a pGL-3 control vector backbone, 5 μg of pGL3-control was digested with *Nco* I and *Xba* I in 50 μl final volume with 2 μl of each enzyme and 5 μl 10X buffer B (nanopure water was used to fill the volume to 50 μl). The digestion reaction was incubated at 37°C for 2 hours, and the whole mixture was run on a 1% agarose gel in 1XTAE. The desired vector backbone fragment was purified using Qiagen's QIAquick gel extraction kit.

The native *Renilla* luciferase gene fragment was cloned into pGL3-control vector using two oligonucleotides, *Nco* I-RL-F and *Xba* I-RL-R, to PCR amplify native *Renilla* luciferase gene using pRL-CMV as the template. The sequence for *Nco* I-RL-F is 5′- CGCTAGCCATGGCTTCGAAAGTTTATGATCC -3′ (SEQ ID NO:238); the sequence for *Xba* I-RL-R is 5′ GGCCAGTAACTCTAGAATTATTGTT-3′ (SEQ ID NO:239). The PCR reaction was carried out as follows:

Reaction mixture (for 100 µl):

DNA template (Plasmid)	1.0 µl (1.0 ng/µl final)
10 X Rec. Buffer	10.0 µl (Stratagene Corp.)
dNTPs (25 mM each)	1.0 µl (final 250 µM)
Primer 1 (10 µM)	2.0 µl (0.2 µM final)
Primer 2 (10 µM)	2.0 µl (0.2 µM final)
Pfu DNA Polymerase	2.0 µl (2.5 U/µl, Stratagene Corp.)
	82.0 µl double distilled water

PCR Reaction: heat 94°C for 2 minutes; (94°C for 20 seconds; 65°C for 1 minute; 72°C for 2 minutes; then 72°C for 5 minutes) x 25 cycles, then incubate on ice. The PCR amplified fragment was cut from a gel, and the DNA purified and stored at –20°C.

To introduce native *Renilla* luciferase gene fragment into pGL3-control vector, 5 µg of the PCR product of the native *Renilla* luciferase gene (RAM-RL-synthetic) was digested with *Nco* I and *Xba* I. The desired *Renilla* luciferase gene fragment was purified and stored at -20°C.

Then 100 ng of insert and 100 ng of pGL3-control vector backbone were digested with restriction enzymes *Nco* I and *Xba* I and ligated together. Then 2 µl of the ligation mixture was transformed into JM109 competent cells. Eight ampicillin resistance clones were picked and their DNA isolated. DNA from each positive clone of pGL3-control-native and pGL3-control-synthetic was purified. The correct sequences for the native gene and the synthetic gene in the vectors were confirmed by DNA sequencing.

To determine whether the synthetic *Renilla* luciferase gene has improved expression in mammalian cells, the gene was cloned into the mammalian expression vector pGL3-control vector under the control of SV40 promoter and SV40 early enhancer (Fig. 13A). The native *Renilla* luciferase gene was also cloned into the pGL-3 control vector so that the expression from synthetic gene and the native gene could be compared. The expression vectors were then transfected into four common mammalian cell lines (CHO, NIH3T3, Hela and CV-1; Table 10), and the expression levels compared between the vectors with the synthetic gene versus the native gene. The amount of DNA used was at two different levels to ascertain that expression from the synthetic gene is consistently increased at different expression levels. The results show a 70-600 fold increase of expression for the synthetic *Renilla* luciferase gene in these cells (Table 10).

Table 10

Enhanced Synthetic *Renilla* Gene Expression

Cell Type	Amount Vector	Fold Expression Increase
CHO	0.2 µg	142
	2.8 µg	145
NIH3T3	0.2 µg	326
	2.0 µg	593
HeLa	0.2 µg	185
	1.0 µg	103
CV-1	0.2 µg	68
	2.0 µg	72

One important advantage of luciferase reporter is its short protein half-life. The enhanced expression could also result from extended protein half-life and, if so, this gives an undesired disadvantage of the new gene. This possibility is ruled out by a cycloheximide chase ("CHX Chase") experiment (Figure 14), which demonstrated that there was no increase of protein half-life resulted from the humanized *Renilla* luciferase gene.

To ensure that the increase in expression is not limited to one expression vector backbone, is promoter specific and/or cell specific, a synthetic *Renilla* gene (Rluc-final) as well as native *Renilla* gene were cloned into different vector backbones and under different promoters (Figure 13B). The synthetic gene always exhibited increased expression compared to its wild-type counterpart (Table 11).

Table 11

Renilla Gene Expression: native v. synthetic (Rluc-final)

Vector	NIH-3T3	HeLa	CHO
pRL-tk, native	3,834.6	922.4	7,671.9
pRL-tk, synthetic	13,252.5	9,040.2	41,743.5
pRL-CMV, native	168,062.2	842,482.5	153,539.5
pRL-CMV, synthetic	2,168,129	8,440,306	2,532,576
pRL-SV40, native	224,224.4	346,787.6	85,323.6
pRL-SV40, synthetic	1,469,588	2,632,510	1,422,830
pRL-null, native	2,853.8	431.7	2,434
pRL-null, synthetic	9,151.17	2,439	28,317.1
pRGL3b, native	12	21.8	17
pRGL3b, synthetic	130.5	212.4	1,094.5
pRGL3-tk, native	27.9	155.5	186.4
pRGL3-tk, synthetic	6,778.2	8,782.5	9,685.9

pRL-tk no intron, native	31.8	165	93.4
pRL-tk no intron, synthetic	6,665.5	6,379	21,433.1

Table 12

Renilla Luciferase Expression in Mammalian Cells

	Percent of control vector		
Vector	CHO cells	NIH3T3 cells	HeLa cells
pRL-control native	100	100	100
pRL-control synthetic	100	100	100
pRL-basic native	4.1	5.6	0.2
pRL-basic synthetic	0.4	0.1	0.0
pRL-promoter native	5.9	7.8	0.6
pRL-promoter synthetic	15.0	9.9	1.1
pRL-enhancer native	42.1	123.9	52.7
pRL-enhancer synthetic	2.6	1.5	5.4

(Vector backbones illustrated in Figure 13A)

With reduced spurious expression the synthetic gene should exhibit less basal level transcription in a promoterless vector. The synthetic and native *Renilla* luciferase genes were cloned into the pGL3-basic vector to compare the basal level of transcription. Because the synthetic gene itself has increased expression efficiency, the activity from the promoterless vector cannot be compared directly to judge the difference in basal transcription, rather, this is taken into consideration by comparing the percentage of activity from the promoterless vector in reference to the control vector (expression from the basic vector divided by the expression in the fully functional expression vector with both promoter and enhancer elements). The data demonstrate that the synthetic *Renilla* luciferase has a lower level of basal transcription than the native gene (Table 12)

It is well known to those skilled in the art that an enhancer can substantially stimulate promoter activity. To test whether the synthetic gene has reduced risk of inappropriate transcriptional characteristics, the native and synthetic gene were introduced into a vector with an enhancer element (pGL3-enhancer vector). Because the synthetic gene has higher expression efficiency, the activity of both cannot be compared directly to compare the level of transcription in the presence of the enhancer, however, this is taken into account by using the percentage of activity from enhancer vector in reference to the control vector (expression in the presence of enhancer divided by the expression in the fully functional expression vector with both promoter and enhancer elements). Such results show that when native gene is present, the enhancer alone is able to stimulate transcription from 42-124% of the control, however, when the native gene is replaced by the synthetic gene in the same vector, the activity only constitutes 1-5% of the value when the same enhancer and a strong SV40 promoter are employed. This clearly demonstrates that synthetic gene has reduced risk of spurious expression (Table 12).

The synthetic *Renilla* gene (Rluc-final) was used in *in vitro* systems to compare translation efficiency with the native gene. In a T7 quick coupled transcription/translation system (Promega Corp., Madison, WI), pRL-null native plasmid (having the native *Renilla* luciferase gene under the control of the T7 promoter) or the same amount of pRL-null-synthetic plasmid (having the synthetic *Renilla* luciferase gene under the control of the T7 promoter) was added to the TNT reaction mixture and luciferase activity measured every 5 minutes up to 60 minutes. Dual Luciferase assay kit (Promega Corp.) was used to measure *Renilla* luciferase activity. The data showed that improved expression was obtained from the synthetic gene (Figure 15A,B). To further evidence the increased translation efficiency of the synthetic gene, RNA was prepared by an *in vitro* transcription system, then purified. pRL-null (native or synthetic) vectors were linearized with *Bam*H I. The DNA was purified by multiple phenol-chloroform extraction followed by ethanol precipitation. An *in vitro* T7 transcription system was employed by prepare RNAs. The DNA template was removed by using RNase-free DNase, and RNA was purified by phenol-chloroform extraction followed by multiple isopropanol precipitations. The same amount of purified RNA, either for the synthetic gene or the native gene, was then added to a rabbit reticulocyte lysate (Figure 15 C, D) or wheat germ lysate (Figure 15 E, F). Again, the synthetic *Renilla* luciferase gene RNA produced more luciferase than the native one. These data suggest that the translation efficiency is improved by the synthetic sequence. To determine why the synthetic gene was highly expressed in wheat germ, plant codon usage was determined. The lowest usage codons in higher plants coincided with those in mammals.

Reporter gene assays are widely used to study transcriptional regulation events. This is often carried out in co-transfection experiments, in which, along with the primary reporter construct containing the testing promoter, a second control reporter under a constitutive promoter is transfected into cells as an internal control to normalize experimental variations including transfection efficiencies between the samples. Control reporter signal, potential promoter cross talk between the control reporter and primary reporter, as well as potential regulation of the control reporter by experimental conditions, are important aspects to consider for selecting a reliable co-reporter vector.

As described above, vector constructs were made by cloning synthetic *Renilla* luciferase gene into different vector backbones under different promoters. All the constructs showed higher expression in the three mammalian cell lines tested (Table 11). Thus, with better expression efficiency, the synthetic *Renilla* luciferase gives out higher signal when transfected into mammalian cells.

Because a higher signal is obtained, less promoter activity is required to achieve the same reporter signal, this reduced risk of promoter interference. CHO cells were transfected with 50 ng pGL3-control (firefly *luc+)* plus one of 5 different amounts of native pRL-TK plasmid (50, 100, 500, 1000, or 2000 ng) or synthetic pRL-TK (5, 10, 50, 100, or 200 ng). To each transfection, pUC19 carrier DNA was added to a total of 3 μg DNA. Shown in Figure 16 is the experiment demonstrating that 10 fold less pRL-TK DNA gives similar or more signal as the native gene, with reduced risk of inhibiting expression from the primary reporter pGL3-control.

Experimental treatment sometimes may activate cryptic sites within the gene and cause induction or suppression of the co-reporter expression, which would compromise its function as co-reporter for normalization of transfection efficiencies. One example is that TPA induces expression of co-reporter vectors harboring the wild-type gene when transfecting MCF-7 cells. 500 ng pRL-TK (native), 5 µg native and synthetic pRG-B, 2.5 µg native and synthetic pRG-TK were transfected per well of MCF-7 cells. 100 ng/well pGL3-control (firefly luc+) was co-transfected with all RL plasmids. Carrier DNA, pUC19, was used to bring the total DNA transfected to 5.1 µg/well. 15.3 µl TransFast Transfection Reagent (Promega Corp., Madison, WI) was added per well. Sixteen hours later, cells were trypsinized, pooled and split into six wells of a 6-well dish and allowed to attach to the well for 8 hours. Three wells were then treated with the 0.2 nM of the tumor promoter, TPA (phorbol-12-myristate-13-acetate, Calbiochem #524400-S), and three wells were mock treated with 20 µl DMSO. Cells were harvested with 0.4 ml Passive Lysis Buffer 24 hours post TPA addition. The results showed that by using the synthetic gene, undesirable change of co-reporter expression by experimental stimuli can be avoided (Table 13). This demonstrates that using synthetic gene can reduce the risk of anomalous expression.

Table 13

TPA Induction

Vector	Rlu	Fold Induction
pRL-tk untreated (native)	184	
pRL-tk TPA treated (native)	812	4.4
pRG-B untreated (native)	1	
pRG-B TPA treated (native)	8	8.0
pRG-B untreated (final)	132	
pRG-B TPA treated (final)	195	1.47
pRG-tk untreated (native)	44	

Vector	Rlu	Fold Induction
pRG-tk TPA treated (native)	192	4.36
pRG-tk untreated (final)	12,816	
pRG-tk TPA treated (final)	11,347	0.88

References

Altschul et al., Nucl. Acids Res., 25, 3389 (1997).

Aota et al., Nucl. Acids Res., 16, 315 (1988).

Boshart et al., Cell, 41, 521 (1985).

Bronstein et al., Cal. Biochem., 219, 169 (1994).

Corpet et al., Nucl. Acids Res., 16, 881 (1988).

deWet et al., Mol. Cell. Biol., 7, 725 (1987).

Dijkema et al., EMBO J., 4, 761 (1985).

Faist and Meyer, Nucl. Acids Res., 20, 26 (1992).

Gorman et al., Proc. Natl. Acad. Sci. USA, 79, 6777 (1982).

Higgins et al., Gene, 73, 237 (1985).

Higgins et al., CABIOS, 5, 151 (1989).

Huang et al., CABIOS, 8, 155 (1992).

Itolcik et al., PNAS, 94, 12410 (1997).

Johnson et al., Mol. Reprod. Devel., 50, 377 (1998).

Jones et al., Mol. Cell. Biol., 17, 6970 (1997).

Karlin and Altschul, Proc. Natl. Acad. Sci. USA, 87, 2264 (1990).

Karlin and Altschul, Proc. Natl. Acad. Sci. USA, 90, 5873 (1993).

Keller et al., J. Cell Biol., 84, 3264 (1987).

Kim et al., Gene, 91, 217 (1990).

Lamb et al., Mol. Reprod. Devel., 51, 218 (1998).

Mariatis et al., Science, 236, 1237 (1987).

Michael et al., EMBO. J., 9, 481 (1990).

Mizushima and Nagata, Nucl. Acids Res., 18, 5322 (1990).

Murray et al., Nucl. Acids Res., 17, 477 (1989).

Myers and Miller, CABIOS, 4, 11 (1988).

Needleman and Wunsen, J. Mol. Biol., 48, 443 (1970).

Pearson and Lipman, Proc. Natl. Acad. Sci. USA, 85, 2444 (1988).

Pearson et al., Meth. Mol. Biol., 24, 307 (1994).

Sharp et al., Nucl. Acids Res., 16, 8207 (1988).

Sharp et al., Nucl. Acids Res., 15, 1281 (1987).

Smith and Waterman, Adv. Appl. Math., 2, 482 (1981).

Stemmer et al., Gene, 164, 49 (1995).

Uetsuki et al., J. Biol. Chem., 264, 5791 (1989).

Voss et al., Trends Biochem. Sci., 11, 287 (1986).

Wada et al., Nucl. Acids Res., 18, 2367 (1990).

Watson et al, eds. Recombinant DNA: A Short Course, Scientific American Books, W. H. Freeman and Company, New York (1983).

Wood, K. Photochemistry and Photobiology, 62, 662 (1995).

Wood, K. Science 244, 700 (1989)

All publications, patents and patent applications are incorporated herein by reference. While in the foregoing specification, this invention has been described in relation to certain preferred embodiments thereof, and many details have been set forth for purposes of illustration, it will be apparent to those skilled in the art that the invention is susceptible to additional embodiments and that certain of the details herein may be varied considerably without departing from the basic principles of the invention.